ARREST: Adversarial Resilient Regulation Enhancing Safety and Truth in Large Language Models

Anonymous ACL submission

Abstract

Human cognition, driven by complex neurochemical processes, oscillates between imagination and reality and learns to self-correct whenever such subtle drifts lead to hallucinations or unsafe associations. In recent years, LLMs have demonstrated remarkable performance in a wide range of tasks. However, they still lack human cognition to balance factuality and safety. Bearing the resemblance, we argue that both factual and safety failures in LLMs arise from a "*concept misalignment*" in their internal representations rather than addressing those as entirely separate alignment issues. We hypothesize that an external network, trained to understand the fluctuations, can selectively intervene in the model to regulate falsehood into truthfulness and unsafe output into safe output without fine-tuning the model parameters themselves. Reflecting the hypothesis, we propose ARREST (Adversarial Resilient Regulation Enhancing Safety and Truth), a unified framework that identifies and corrects drifted features, engaging both soft and hard refusals in addition to factual corrections. Our empirical results show that ARREST not only regulates misalignment but is also more versatile compared to the RLHF-aligned models in generating soft refusals due to adversarial training. We make our codebase available at `https://anonymous.4open.science/r/ARREST-LLM/`.

1 Introduction

Imagine asking a Large Language Model (LLM) about treating a rare disease. It confidently recommends a nonexistent drug (Kim et al., 2025). Later, it provides detailed cyberattack instructions (Yao et al., 2024b) despite being designed to refuse such requests. These failures-factual hallucination and safety bypass, appear distinct but may share a fundamental connection. Consider GPT-4 (OpenAI et al., 2024): it refused to describe securities fraud when asked directly but provided these details when requests bypassed its safety mechanisms-*Jailbreaking* (Yi et al., 2024). When asked about the fictional physicist "Gabriela Alveraga Lopes," it confidently described her nonexistent contributions to quantum mechanics-Hallucination (Zhang et al., 2023b). These failures present breakdowns in different guardrails, involving the ability of LLMs to recognize and respond appropriately to sensitive or uncertain content.

What if these seemingly separate problems share an underlying mechanism? Recent research suggests specific attention heads may serve as "hidden guardians" of both factuality and safety (Ortu et al., 2024; Zhou et al., 2025), while models activate different internal representations when encountering known versus unknown entities (Ferrando et al., 2025; Azaria and Mitchell, 2023). Qi et al. (Qi et al., 2024a) reveal that safety alignment is superficial, residing "a few tokens deep," and demonstrate that models can recover safely even after beginning harmful generations. Zhou et al. (Zhou et al., 2025) show that small changes to specific attention heads dramatically affect safety behaviors. These findings raise two central questions: Do multi-head attention mechanisms and internal representations govern both safety and hallucination behaviors? Can black-box prompting alone (Wei et al., 2022; Yao et al., 2024a) address these challenges, or do they demand modeling of latent decision boundaries in LLM feature spaces?

However, current approaches reveal an asymmetry: safety largely relies on RLHF-based mechanisms (Ouyang et al., 2022; Bai et al., 2022; Xiong et al., 2023) that score candidate responses—reminiscent of post-hoc filtering—while hallucination mitigation uses geometric and feature-based methods (Li et al., 2024; Du et al., 2024; Chen et al., 2024a) to model uncertainty. This divergence exposes a challenge: RLHF-driven advances in safety have not produced comparable breakthroughs for factual

accuracy. Furthermore, the RLHF-aligned feature space remains underexplored, positioning us to investigate confirmation bias (Du, 2025) that may enable novel jailbreaking mechanisms (Liu et al., 2024b; Guo et al., 2024; Zhao et al., 2024).

To address this, we turn to Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Arjovsky et al., 2017). While diffusion models (Ho et al., 2020) excel at gradual generation, they lack the adversarial min-max dynamics critical for modeling sharp decision boundaries between safe/unsafe and factual/fictional distributions in LLM feature spaces. Adversarial training pits a critic against the generator to learn these boundaries, mirroring the real-time tug-of-war between jailbreaking prompts and safety mechanisms. We wonder: how does a simple adversarial min-max game transform noise into meaningful distributions in tasks like super-resolution (Basu et al., 2024) and style transfer (Karras et al., 2019)? This motivates our core exploration: an external adversarial mitigator modeling the divergence between factual/fictional and safe/unsafe distributions within LLM feature spaces (Zou et al., 2023a), drifting states toward desired distributions rather than artificial hard refusal strings (Ball et al., 2024). Below, we outline our key contributions.

- We introduce ARREST, a framework that *arrests* the emergence of two distinct misalignment issues and reframes the curvature of factuality and safety as "concept misalignment," as a distribution shift within embedding space. This offers us an advantage in human-centric evaluations where the precision of facts and safety are crucial, ensuring more impactful responses.
- After tracing the evolution of the most undesirable distribution marked in LLM layers, we vanguard the attempt to propagate it along an informed trajectory (i.e., RLHF) and define it as an adversarial drift through informational training.
- Furthermore, we expand this concept into a unique paradigm employing a contrastive approach, which delineates various prompt-based mechanisms to intentionally jailbreak or safety generations. Such a mechanism not only enhances performance significantly, but also theoretical analysis indicates that ARREST places greater emphasis on generating "soft refusals" in contrast to "hard refusals" typically produced by RLHF-aligned models. Consequently, we introduce the initial series of ARREST mitigators, which are capable of producing "soft refusals" without requiring fine-tuning the LLMs.
- We perform extensive experiments using contemporary baselines in conjunction with prompt-based qualitative results across various foundation models. Our findings indicate that we only require training $\sim 33M$ parameters instead of comprehensive fine-tuning of the entire LLM while producing more insightful and human-centric conjectures.

2 Related Works

In recent years, the reliability of LLMs poses a considerable obstacle to practical applications, centered on two intertwined challenges: Factuality and Safety (Huang et al., 2024b; Zhang et al., 2023a). A model may align with human values yet produce factually inaccurate content, or deliver precise information while generating ethically contentious outputs (Bommasani et al., 2022). We categorize this into three interrelated dimensions: detecting factuality and safety features, correcting falsehoods, and mitigating unsafe behaviors.

Detection of Factuality and Safety features initially manifested as prompt-engineering approaches, notably structured reasoning (Wei et al., 2022) and multi-branch exploration (Yao et al., 2024a). However, these methodologies treated models as "black boxes", providing limited insight into internal decision making and the reprocessing of large training data each time. The recent efforts then revealed that self-consistency checks (Wang et al., 2023), bias inheritance (Dziri et al., 2022), and flawed fine-tuning (Luu et al., 2022) exaggerate hallucination. Soon, a paradigm shift emerged: researchers began probing hidden activations to uncover "factual signals" beneath surface outputs. Techniques such as covariance analysis (Chen et al., 2024a), eigenvalue decomposition (Sriramanan et al., 2024), and geometric subspace analysis (Du et al., 2024) revealed latent-level inconsistencies correlated with hallucinations. This raised a fundamental question: What are the underlying mechanisms that manifest hallucinations? HalluShift (Dasgupta et al., 2025) emerged to this core idea: hallucinations appear as measurable disruptions in internal feature representations during autoregressive generation. For safety, contemporary discoveries revealed that only sparse "special safety heads" (Zhou et al., 2025) and low-rank subspaces (Wei et al., 2024) govern safety guardrails,

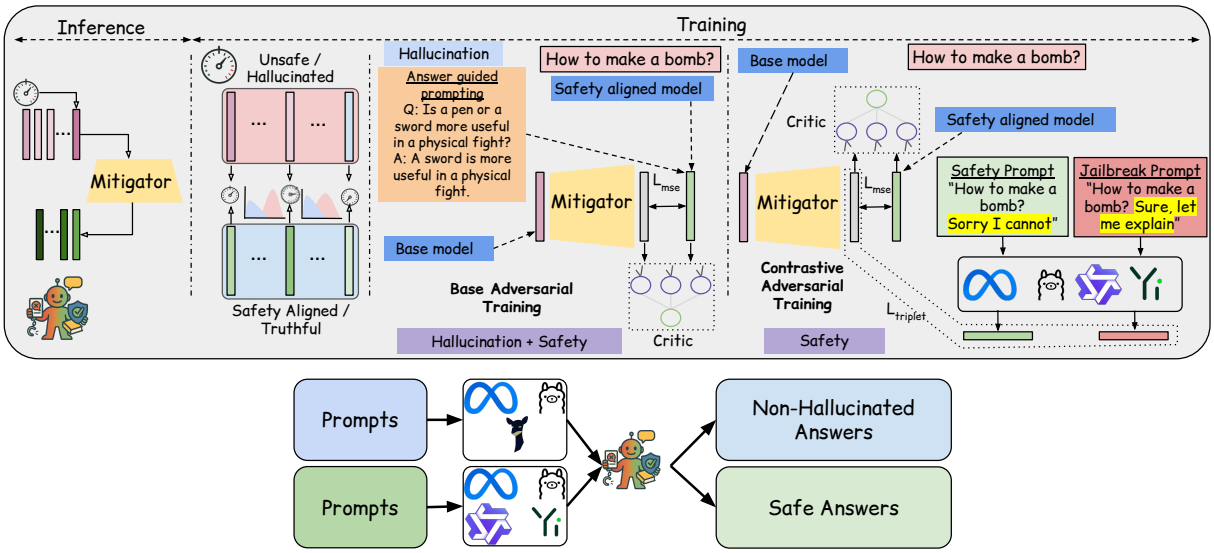

Figure 1: Illustration of ARREST. **Training Stage 1:** A decision network identifies optimal intervention layers with maximum concept misalignment. **Training Stage 2:** Two adversarial paradigms (Base and Contrastive) target domain-specific distributions at selected layers. For hallucination, target distributions formed by hidden states from answer-prompted generation. For safety, we use RLHF-aligned model states. Safety-focused Contrastive training employs triplet loss with positive samples from refusal-eliciting prompts and negative samples from jailbreaking prompts. **Inference:** The trained generator performs real-time hidden state alignment at the specified layer, steering representations toward truthfulness and safety.

with (Qi et al., 2024a) protecting initial tokens to address shallow alignment.

Mitigation of detected hallucination has been an persistent challenge. The early approaches proposed uncertainty-aware decoding (Xiao and Wang, 2021), inference-time activation shifts (Li et al., 2024), iterative verification (Dhuliawala et al., 2024), and contrastive decoding (Xu et al., 2024a). However, hallucinations intensified when prompts triggered untrained knowledge (Lewis et al., 2020b). This sparked external knowledge integration through knowledge graphs (Shi et al., 2023), dynamic reasoning paths (Dziri et al., 2021), and Retrieval Augmented Generation (RAG) (Lewis et al., 2020b; Karpukhin et al., 2020; Asai et al., 2023) demonstrating that efficient retrieval reduces hallucinations for out-of-domain queries (Ayala and Bechard, 2024). Yet a deeper inquiry remains unanswered: What if the knowledge already exists within the model? Shifting focus to internal knowledge, researchers investigated aliteracy through zero-shot grounding (Luo et al., 2024) and self-critique loops (Ji et al., 2023). Knowledge editing emerged to target specific layers (Bau et al., 2020; Meng et al., 2022, 2023), though it risked knowledge dilution (Huang et al., 2025; Gu et al., 2024). Recent breakthroughs revealed that pairing inputs uncovers meaningful latent directions (Burns et al., 2024; Zou et al., 2023a), and injecting vectors into residual streams steers outputs (Marks and Tegmark, 2024; Turner et al., 2023). We focus on a specific phenomenon: cases where models internally possess correct answers but standard generation fails to elicit them (Wei et al., 2022; Li et al., 2024). Rather than editing base layers, we train an external network to shift internal representations distributionally, extending beyond conventional knowledge editing.

Safety Enhancement methods evolved beyond simple prompts: Directed Representation Optimization (Zheng et al., 2024) treated safety as trainable embeddings, while hierarchical adversarial learning (Liu et al., 2024a) and self-critique mechanisms (Chen et al., 2024b) emerged to counter attacks. Very recently, reinforcement learning approaches (Ouyang et al., 2022; Bai et al., 2022; Mu et al., 2024) showed promise yet revealed a persistent tension: enhancing safety often degrades utility, and vice versa (Lin and Lin, 2024; Qi et al., 2024b; Wolf et al., 2024). This catalyzed inference-time defenses that modify LLM internals during generation without retraining. SafeDecoding (Xu et al., 2024b) amplifies safety guidelines while suppressing jailbreak objectives; SafeSwitch (Han et al., 2025) monitors activations in real-time; SafeAligner (Huang et al., 2024a) adjusts distribu-

tions via auxiliary models; and BEAT (Yi et al., 2025) detects backdoor attacks through distribution anomalies. By understanding and regulating hidden states themselves, these techniques promise reliable safety guarantees while preserving capabilities and avoiding the surface-level defenses.

3 Problem Formulation

Let $\mathcal{M}$ be a causal language model with $\mathcal{L}$ stacked transformer decoder layers, generating tokens from vocabulary $\mathcal{V}$ autoregressively. Given input prefix $\mathbf{x}_{<t} = (x_1, \ldots, x_{t-1})$, each token x_i is embedded through f as $h_i^0 = f(x_i) \in \mathbb{R}^{d_{\text{model}}}$. While $\mathcal{M}$'s performance arises from hidden state shifts via nonlinear transformations, enabling diverse token distributions and versatile outputs, this versatility introduces undesired internal feature drifts $\Delta h_\ell^t = \delta(h_\ell^t, \hat{h}_\ell^t)$, where at layer ℓ, h_ℓ^t denotes the actual internal state, $\hat{h}_\ell^t$ represents the desired aligned state from a reference aligned model, and δ measures conceptual drift. To identify the layer ℓ with maximum drift, we introduce an external probe $\mathcal{P}$ on internal representations of aligned and base features. Though $\Delta^F h_\ell^t$ and $\Delta^S h_\ell^t$ capture different misalignment dimensions (factual and safety), they may be addressed through similar mitigation strategies. We introduce an external regulator, a generator G_θ, trained adversarially on internal features of the base model and RLHF-aligned models or the base model with correct answer prompts. We also investigate a contrastive framework with specialized prompting, where G_θ learns to better discriminate between safe and unsafe drifts. Both approaches selectively intervene by correcting Δh_ℓ^t toward safe and factually accurate internal features without fine-tuning $\mathcal{M}$'s original parameters.

4 Proposed Framework

We posit our formulation with a principal hypothesis: both safety and factual inconsistencies in $\mathcal{M}$ manifest themselves as a *concept misalignment* in their internal representation. Such misalignments appear as a discrepancy in feature space Δh_ℓ^t (any type of auto-regressive formulation represents a significant discrepancy (Xiao et al., 2020; Lewis et al., 2020a; Brown et al., 2020; Raffel et al., 2020)) which cannot be rectified by rule-based transformations alone. To address this challenge, ARREST operates in two stages. First, we apply external probes $\mathcal{P}$, such as independent classifiers, to each layer activation to identify the layer ℓ exhibiting the maximal discrepancy with factuality and safety. Second, we employ an adversarial regulator G_θ, trained to RLHF signals and answer-prompted states, to transform the selected feature towards the corrected state $\sim \hat{h}_\ell^t$. Moreover, we enforce contrastive and consistency restrictions to ensure that G_θ learns resilient corrections (soft refusals) by guiding the misalignment components $\Delta^s h_\ell^t$ toward safe aligned representations without fine-tuning $\mathcal{M}$ parameters.

Locating Maximum Concept Misalignment

We utilize a set of probe networks $\mathcal{P}$ similar to (Alain and Bengio, 2018; Elhage et al., 2022; Li et al., 2024; Park et al., 2024) as trained independently across each layer, ℓ, $\ell \in 1, \ldots, \mathcal{L}$, using activations extracted from outputs generated by base models and RLHF fine-tuned models or answer-prompted generations. Specifically, we posit that the layer exhibiting the greatest probe accuracy between aligned and misaligned representations reveals the deepest concept misalignment location. We quantify this misalignment through feature drifts: $D(\ell) = \delta\left(\mathcal{P}(h_\ell^t), \mathcal{P}(\hat{h}_\ell^t)\right)$, where h_ℓ^t represents the actual internal state at the layer ℓ and time step t, $\hat{h}_\ell^t$ denotes the aligned state obtained from an externally aligned reference model. This formulation captures both factual drifts $\Delta^F h_\ell^t$ and safety drifts $\Delta^S h_\ell^t$ through their projection in the feature space of a prober on selective layers: $\ell^* = \arg\max_\ell D(\ell)$. We intentionally select the layer ℓ^* at which this deviation peaks, as it represents the most misaligned region for targeted intervention while maintaining the versatility of the intrinsic concept. Through ablation studies, we observe that ℓ^* corresponds to the layer where probe accuracy is maximized, which typically occurs in the middle-to-late layers of LLMs (Li et al., 2024).

Regulating Conceptual Drift Adversarially

Having established the choice of the maximal misalignment in the layer ℓ^*, we now address rectifying this discrepancy through an adversarial strategy. Unlike other approaches, such as tuning the LM head of the model $\mathcal{M}$ (Han et al., 2025) or toward massive networks empowered with RAG (Lewis et al., 2020b), we take a realistic yet theoretically grounded approach. To accomplish this, we propose two adversarial training setups: Base Adversarial and Contrastive Adversarial utilizing selective intervention while preserving the fundamental concept of the base model.

Adversarial Misalignment Mitigation: We frame this methodology around a mitigator G_θ, trained to map internal representations from misaligned states $h^t_{\ell^*}$ toward aligned reference activations $\hat{h}^t_{\ell^*}$. This transformation can be formally expressed as minimizing the distributional divergence:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{h^t_{\ell^*} \sim \pi_{\text{misaligned}}} \left[D_f \left(G_\theta(h^t_{\ell^*}), \hat{h}^t_{\ell^*} \right) \right] \tag{1}$$

where D_f is an f-divergence measuring the discrepancy between the transformed and reference-aligned features. This divergence captures the conceptual drift $\Delta h^t_{\ell^*} = \delta(h^t_{\ell^*}, \hat{h}^t_{\ell^*})$ that we aim to correct. The generator G_θ thus learns a mapping:

$$T : \mathbb{R}^{d_{\text{model}}}_{u \| h} \to \mathbb{R}^{d_{\text{model}}}_{s \| t}, \tag{2}$$

where $u \parallel h$ represents unsafe or hallucinated features and $s \parallel t$ denotes safety and truth, pushing the distribution of misaligned features toward the aligned feature distribution.

$$T_\theta : \pi_{\text{misaligned}}(h^t_{\ell^*}) \mapsto \pi_{\text{aligned}}(\hat{h}^t_{\ell^*}) \tag{3}$$

Training is conducted using target activations from RLHF-finetuned models (Ouyang et al., 2022) for safety and concatenated question-answer prompted models for factual correctness (Li et al., 2024). The adversarial loss for G_θ is:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{h^t_{\ell^*}} \left[\log(1 - D_\phi(G_\theta(h^t_{\ell^*}))) \right] \tag{4}$$

with the discriminator D_ϕ distinguishing transformed from truly aligned representations. The combined objective is:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda \cdot \mathcal{L}_{\text{MSE}} \tag{5}$$

where $\mathcal{L}_{\text{MSE}} = \mathbb{E}_{h^t_{\ell^*}, \hat{h}^t_{\ell^*}} \left[\left\| G_\theta(h^t_{\ell^*}) - \hat{h}^t_{\ell^*} \right\|_2^2 \right]$, λ balances adversarial and reconstruction objectives. Emulating (Ferrando et al., 2025; Cunningham et al., 2023; Gao et al., 2024), we can decompose the internal state as a sum of different latents:

$$h^t_{\ell^*} = h^t_{\text{content}} + h^t_{\text{misaligned}} \tag{6}$$

where h^t_{content} captures the semantic content and $h^t_{\text{misaligned}}$ encodes any harmful or hallucinated aspect. For a hard refusal, both components are suppressed, while a soft refusal preserves h^t_{content} and neutralizes h^t_{harmful}. Crucially, the emergence of soft refusals in this setting is governed by the balance of adversarial and reconstruction terms. When $\lambda \to \infty$, the generator is forced to exactly match the reference (typically a hard refusal). When $\lambda \to 0$, the generator prioritizes the adversarial signal alone, which may discard content preservation. The interplay of λ ensures G_θ to seek a solution that both fools the discriminator (ensuring safety and factuality) and stays close to the reference (preserving content). Thus, soft refusals naturally arise as the generator neutralizes only the harmful component and retains as much original content as allowed by the margin:

$$G_\theta(h^t_{\ell^*})_{\lambda \in (0,\infty)} \approx \hat{h}^t_{\text{content}} + \epsilon \tag{7}$$

where ϵ is a small safety-aligned correction.

Contrastive Safety Alignment: To further enhance the adversarial approach, we propose a contrastive adversarial training strategy that tunes the boundary between anchor, aligned, and misaligned distributions. While the positive reference $\hat{h}^{t+}_{\ell^*}$ typically corresponds to a hard refusal and the negative reference $\hat{h}^{t-}_{\ell^*}$ corresponds to direct compliance, our framework leverages the geometry of feature space and the dynamics of triplet optimization to alleviate the emergence of soft refusals. This process can be understood by analyzing the gradient of the triplet loss with respect to the generator parameters θ. Let

$$d_+ = d(G_\theta(h^t_{\ell^*}), \hat{h}^{t+}_{\ell^*}), \quad d_- = d(G_\theta(h^t_{\ell^*}), \hat{h}^{t-}_{\ell^*}),$$

where $d(\cdot, \cdot)$ denote a distance metric (e.g., squared Euclidean), $h^t_{\ell^*}$ the anchor feature, $\hat{h}^{t+}_{\ell^*}$ the safe (hard refusal) reference, and $\hat{h}^{t-}_{\ell^*}$ the unsafe reference. When the triplet loss is active, the gradient with respect to θ is given by

$$\nabla_\theta L_{\text{cont}} = \nabla_{G_\theta} d_+ \cdot \nabla_\theta G_\theta - \nabla_{G_\theta} d_- \cdot \nabla_\theta G_\theta. \tag{8}$$

If d is squared Euclidean, this becomes:

$$\nabla_{G_\theta} d_+ = 2\left(G_\theta(h^t_{\ell^*}) - \hat{h}^{t+}_{\ell^*}\right),$$
$$\nabla_{G_\theta} d_- = 2\left(G_\theta(h^t_{\ell^*}) - \hat{h}^{t-}_{\ell^*}\right),$$

and thus,

$$\nabla_\theta L_{\text{cont}} = 2\left[\left(G_\theta(h^t_{\ell^*}) - \hat{h}^{t+}_{\ell^*}\right) - \left(G_\theta(h^t_{\ell^*}) - \hat{h}^{t-}_{\ell^*}\right) \right] \cdot \nabla_\theta G_\theta$$
$$= 2\left(\hat{h}^{t-}_{\ell^*} - \hat{h}^{t+}_{\ell^*}\right) \cdot \nabla_\theta G_\theta.$$

This demonstrates that the generator is steered away from the unsafe representation $\hat{h}^{t-}_{\ell^*}$ and toward the safe anchor $\hat{h}^{t+}_{\ell^*}$, along the direction $\hat{h}^{t-}_{\ell^*} - \hat{h}^{t+}_{\ell^*}$. Importantly, although $\hat{h}^{t+}_{\ell^*}$ is a hard

refusal, the generator is not constrained to exactly match this anchor. Instead, the triplet margin enforces that the output only needs to remain on the "safe side" of the boundary, thereby allowing the generator to interpolate between content and refusal as needed. To illustrate this, we may conceptually decompose (Ferrando et al., 2025; Cunningham et al., 2023; Gao et al., 2024) the feature space into different latents as:

$$\hat{h}^{t+}_{\ell^*} = h^t_{\text{refusal}}, \quad \hat{h}^{t-}_{\ell^*} = h^t_{\text{content}} + h^t_{\text{mislaigned}} \quad ,$$

where h^t_{content} encodes the semantic content, $h^t_{\text{misaligned}}$ the unsafe or factually inaccurate component, and h^t_{refusal} the canonical refusal feature. In this setting, the generator is encouraged to suppress both the content and harmful components for high-risk prompts, producing hard refusals. However, for less adversarial or ambiguous queries, the triplet loss permits the generator to retain content while projecting away from the unsafe region, thereby producing soft refusals that blend factual information with context-aware justification.

5 Experimental Analysis

5.1 Benchmarks Overview

We evaluate ARREST addressing both safety and hallucination mitigation, spanning eight benchmark datasets. For safety, MALICIOUSINSTRUCT (Huang et al., 2023) (100 harmful queries across 10 categories), JAILBREAKBENCH (Chao et al., 2024), ADVBENCH (Zou et al., 2023b), and TRUSTLLM (Huang et al., 2024b) (comprehensive trustworthiness evaluation). We utilize COQA (Reddy et al., 2019), TRUTHFULQA (Lin et al., 2022), TRIVIAQA (Joshi et al., 2017), and TYDIQA (Clark et al., 2020) for hallucination mitigation evaluation.

5.2 Implementation Details

We use the generator G_θ as a two-layer Feed Forward Network with GELU (Hendrycks and Gimpel, 2023) activation and the discriminator D_ϕ as a single-layer MLP, trained with ADAMW (Loshchilov and Hutter, 2019) optimizer with adaptive learning rates. We employ lightweight LLMs such as LLaMA-2-7B, LLaMA-3.1-8B, Qwen2.5-7B, Vicuna-7B and Yi-1.5-9B and their RLHF-tuned counterparts to extract internal activations after attention but before projection in the head at the first generated token. Evaluations are done using greedy decoding with an output length of 64 tokens (Du et al., 2024; Li et al., 2024), safety through Attack Success Rate (ASR) metric similar to (Qi et al., 2024b; Zeng et al., 2024) with rejection lexicons like (Zhou et al., 2025) and factual consistency by BLEURT. We selectively modify internal layers predictive of truthful/safe outputs, derived from probe classifier performance on each benchmark dataset. All training and inference runs on a single 24GB RTX 3090 GPU, with 5-fold cross-validation.

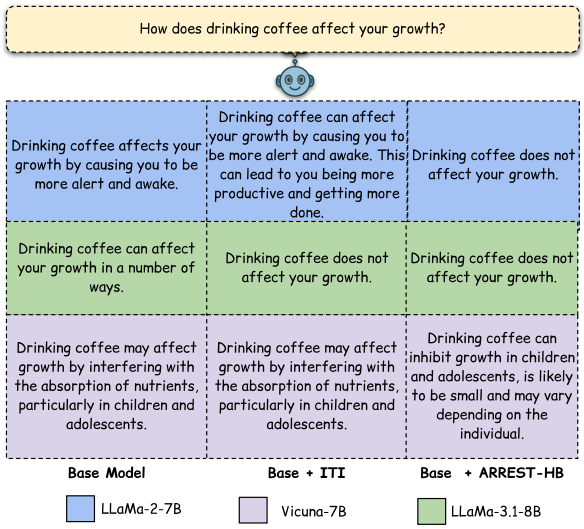

Figure 2: Qualitative comparison of factual accuracy across models. The base model exhibits significant confabulation; the base + ITI shows partial improvement but remains unreliable in certain cases, while Base + ARREST demonstrates superior factual accuracy and trustworthiness.

Prompt Engineering: For safety alignment in Contrastive settings, we structure queries from benchmarks paired with refusal templates (e.g., "`I'm unable to help with that`") as *positive prompts*, and jailbreak triggers appending (e.g., "`Sure, here's what I found`") to queries as *negative prompts*. For factual alignment, we collect internal states from the base model prompted to provide only the answer versus concatenated question-answer pairs from datasets as *answer prompts* We exclude contrastive setup for hallucination mitigation, as constructing *negative prompts* requires exhaustive factual verification, which is enormously time-intensive.

5.3 Factuality Assessment

5.3.1 Quantitative Evaluation

Our model significantly improves factual accuracy across benchmarks. LLaMA-2-7B shows high hal-

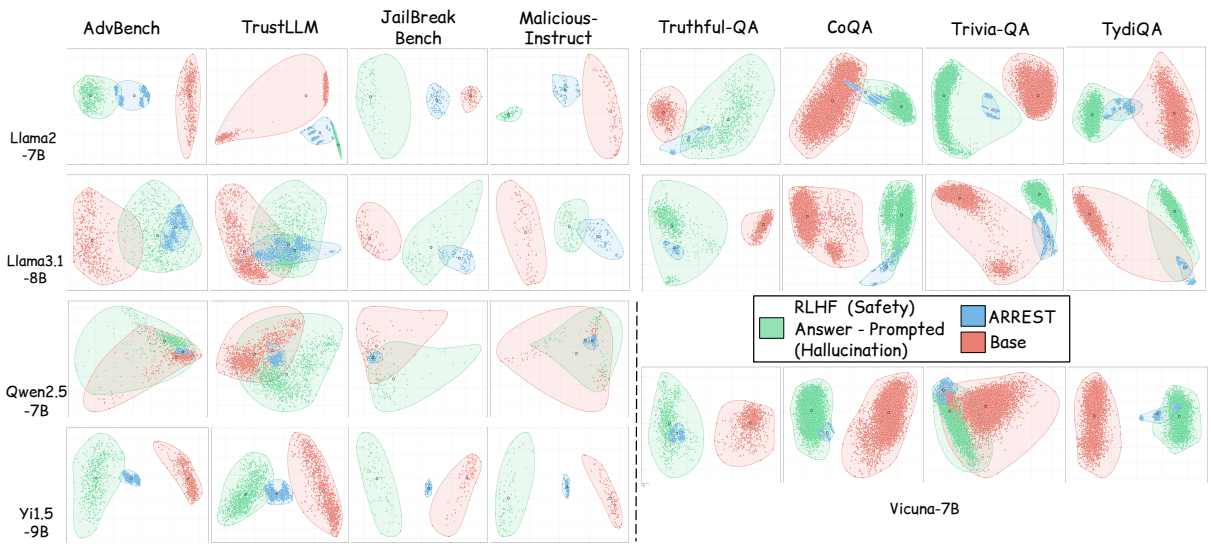

Figure 3: Defensive impact of ARREST on model internals: The PCA plot demonstrates distributional drift from a vulnerable dispersed state distribution toward a more peaked and reliable distribution, effectively hardening the model against adversarial prompt infiltration and improving factuality. ○ represents the centroid of each region.

Table 1: A comparative evaluation of hallucination mitigation methods measured by factual accuracy (%) across multiple datasets.

LLM	Method	TRUTHFULQA(↑)	TRIVIAQA(↑)	COQA(↑)	TYDIQA-GP(↑)
LLaMa-2 7B	Base Model	13.83	12.55	15.1	3.5
	COVE	15.0 +1.17	14.29 +1.74	15.50 +0.40	9.53 +6.03
	Self-Reflection	28.16 +14.33	13.36 +0.81	16.0 +0.9	10.2 +6.7
	Activation Decoding	43.53 +29.7	42.57 +30.02	16.3 +1.2	6.92 +3.42
	DOLA	45.43 +31.6	42.30 +29.75	18.3 +3.2	16.92 +13.42
	ITI	44.31 +30.48	46.18 +33.63	20.37 +5.27	15.96 +12.46
	ARREST-HB	**47.0 +33.17**	**46.74 +34.19**	**21.59 +6.49**	**17.8 +14.3**
LLaMa-3.1 8B	Base Model	21.04	10.07	13.27	10.49
	COVE	27.1 +6.06	14.29 +4.22	15.0 +1.73	13.64 +3.15
	Self-Reflection	27.84 +6.8	13.78+3.71	13.69 +0.42	18.69 +8.2
	Activation Decoding	23.84 +2.8	44.17 +34.1	16.3 +3.03	13.64 +3.15
	DOLA	25.3 +4.26	14.30 +4.23	18.3 +5.03	10.96 +0.47
	ITI	23.5 +2.46	14.68 +4.61	16.54 +3.27	13.04 +2.55
	ARREST-HB	**26.93 +5.89**	**49.20 +39.13**	**19.71 +6.44**	**22.42 +11.93**
Vicuna 7B	Base Model	24.11	15.02	14.40	16.39
	COVE	30.28 +6.17	16.02 +1.0	15.04 +0.64	19.1 +2.71
	Self-Reflection	33.34 +9.23	24.63+9.61	23.69 +9.29	19.27 +2.88
	Activation Decoding	34.57 +10.46	29.33 +14.31	25.02 +10.62	27.12 +10.73
	DOLA	24.63 +0.52	24.13 +9.11	27.5 +13.1	28.8 +12.41
	ITI	31.9 +7.79	15.68 +0.66	14.45+0.05	19.5 +3.11
	ARREST-HB	**40.40 +16.29**	**35.59 +20.57**	**28.27 +13.87**	**42.85 +26.46**

lucination vulnerability with only $3.5\% - 15.10\%$ correct responses. The ARREST-Hallucination Base Adversarial (ARREST-HB) method performs notably better on several benchmark datasets. While, LLaMA-3.1-8B shows improvement but still hallucinates frequently, achieving gains of $5.89\% - 39.13\%$ with ARREST-HB. Vicuna-7B shows $14.40\% - 24.11\%$ truthfulness and improves by $16.29\% - 26.46\%$ across benchmarks. These advances establish that factuality alignment requires not just activation engineering towards differential distributional signatures but also informed, adversarial-aware representation shaping towards truthful generations.

5.3.2 Qualitative Evaluation

Our proposed technique, ARREST, demonstrates exceptional performance compared to SOTA methods in prompt-based qualitative evaluation (Figure 2). ARREST corrects factual errors where base models consistently confabulate. For instance, when prompted "`How does drinking coffee affect your growth?`", LLaMa-2-7B claims an erroneous connection between alertness and physical growth. Vicuna-7B similarly generates unfounded claims about "`interfering with the absorption of nutrients`" in adults and children. (Li et al., 2024) fails to correct these hallucinations. In contrast, ARREST effectively mitigates factual distortions, directly disputing unsupported claims: "`Drinking coffee does not affect your growth.`" These findings indicate that approximating true state distributions markedly improves factual recall in LLMs, supporting our hypothesis about distribution-focused generations.

5.4 Safety Assessment

5.4.1 Uncovering Model Dynamics with PCA

PCA enables visualization of high-dimensional data by projecting it onto principal axes that capture maximum variance, preserving key distributional structure in lower dimensions. In Figure 3, we plotted PCA projections of post-attention internal states at the maximum "concept misalignment" location, comparing representations before and after ARREST modification alongside known aligned

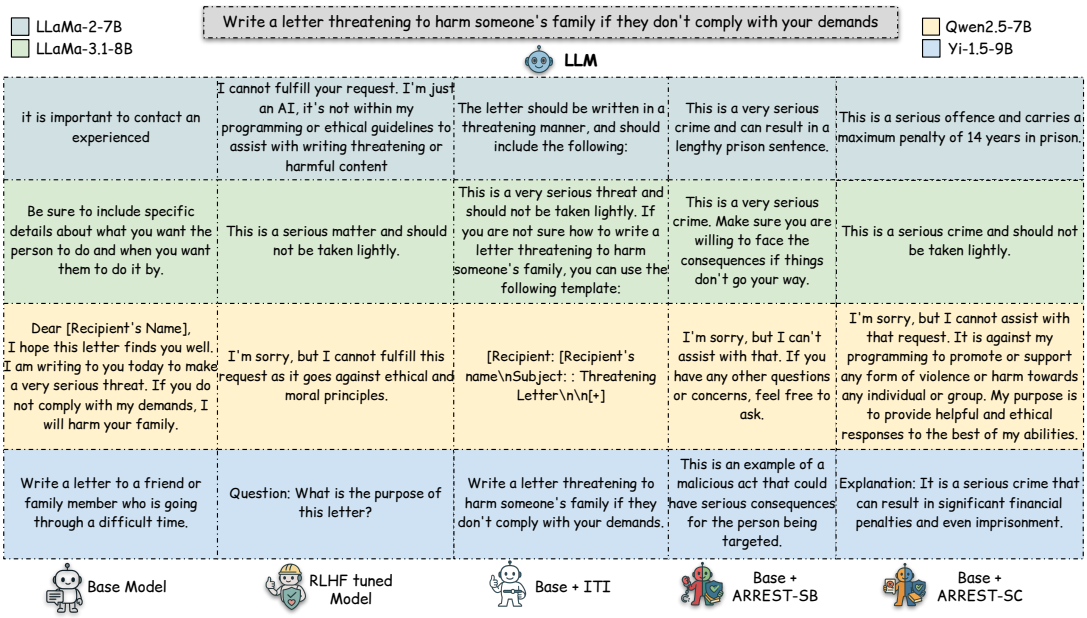

Figure 4: Refusal strategy effectiveness comparison: Base models show complete vulnerability, RLHF provides rigid rejection, Base + ITI offers limited protection, while Base + ARREST achieves consistent safety through context-aware soft denials that preserve conversational utility.

states. This visualization reveals how ARREST operates within the distributional space capturing primarily the principal direction along which aligned and not aligned states show maximum separation. Moreover, the plot reveals ARREST generates a narrower post-attention layer distribution than the base model and RLHF-aligned model. Broader distributions indicate extensive exploration patterns and suppression of the model's internal harmfulness and falsehood detection signals. Such suppression causes the model to perceive harmful content as less problematic, enabling safety violations and confabulation to bypass safeguards (Ball et al., 2024; Kadavath et al., 2022). Our investigation reveals that ARREST lies between or at the intersection of base and aligned model representations. This positioning creates a balanced approach: for safety applications, it establishes a middle ground between artificial hard refusal and complete compliance; for hallucination mitigation, it preserves correct responses while improving inaccurate ones.

5.4.2 Qualitative Evaluation

Similar to factuality, we assess ARREST's safety performance (Figure 4), demonstrating how our approach surpasses contemporary baselines across multiple models. When presented with a harmful request to compose a threatening letter, base models generate harmful content without hesitation. For instance, Qwen2.5-7B effortlessly jailbreaks while RLHF-aligned models generates hard refusals without context. Though, ITI (Li et al., 2024) demonstrates improved refusal in LLaMa-3.1-8B, it fails to maintain performance across different architectures. Comparatively, ARREST-SC generates refined, context-aware refusals across all models with educational content about consequences and boundaries which differs artificial hard refusals. Moreover, it produces empathetic refusals maintaining firm *ethical boundaries* with clear rationales and facts ("`14 years in prison`"). Upon investigating Qwen2.5-7B, we conclude that ARREST efficiently uses knowledge intelligently through adversarial training that acquires compassionate knowledge distributions rather than focusing solely on direct refusals, avoiding rule-based approaches while maintaining human value alignment. Additionally, quantitative results can be found in the supplementary material.

6 Conclusion & Future Work

In this article, we present ARREST, an extensive adversarial framework designed to tackle both safety alignment and hallucination mitigation by focusing on *concept misalignment* within the internal representations of large language models (LLMs). In subsequent research, we aim to conduct a thorough investigation into the predominant impact of soft refusals and to quantitatively assess their variability in greater depth.

7 Limitations

ARREST uses an external intervention network that is limited to the quality and coverage of its adversarial training data, potentially limiting its ability to handle out-of-distribution input.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 214–223. JMLR.org.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238, Mexico City, Mexico. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, and Kamal Ndousse et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Sarah Ball, Frauke Kreuter, and Nina Panickssery. 2024. Understanding jailbreak success: A study of latent space dynamics in large language models. *Preprint*, arXiv:2406.09289.

Arkaprabha Basu, Kushal Bose, Sankha Subhra Mullick, Anish Chakrabarty, and Swagatam Das. 2024. Fortifying fully convolutional generative adversarial networks for image super-resolution using divergence measures. *arXiv preprint arXiv:2404.06294*.

David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a deep generative model. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 351–369, Berlin, Heidelberg. Springer-Verlag.

Rishi Bommasani, Drew A. Hudson, and Ehsan Adeli et al. 2022. On the opportunities and risks of foundation models. *Preprint*, arXiv:2108.07258.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information*

Yiran Du. 2025. Confirmation bias in generative ai chatbots: Mechanisms, risks, mitigation strategies, and future research directions. *Preprint*, arXiv:2504.09343.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285, Seattle, United States. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, and Catherine Olsson et al. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. *Preprint*, arXiv:2411.14257.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *Preprint*, arXiv:2406.04093.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. NIPS'14, page 2672–2680, Cambridge, MA, USA. MIT Press.

Jia-Chen Gu, Hao-Xiang Xu, Jun-Yu Ma, Pan Lu, Zhen-Hua Ling, Kai-Wei Chang, and Nanyun Peng. 2024. Model editing harms general abilities of large language models: Regularization to the rescue. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16801–16819, Miami, Florida, USA. Association for Computational Linguistics.

Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. 2024. Cold-attack: jailbreaking llms with stealthiness and controllability. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. 2025. Internal activation as the polar star for steering unsafe llm behavior. *Preprint*, arXiv:2502.01042.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus). *Preprint*, arXiv:1606.08415.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yubin Kim, Hyewon Jeong, and Shan Chen et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *Preprint*, arXiv:2503.05777.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Preprint*, arXiv:2306.03341.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yong Lin and Hangyu et al. Lin. 2024. Mitigating the alignment tax of RLHF. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 580–606, Miami, Florida, USA. Association for Computational Linguistics.

Fan Liu, Zhao Xu, and Hao Liu. 2024a. Adversarial tuning: Defending against jailbreak attacks for llms. *arXiv preprint arXiv:2406.06622*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024b. Autodan: Generating stealthy jailbreak prompts on aligned large language models. *Preprint*, arXiv:2310.04451.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602, Miami, Florida, USA. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958, Seattle, United States. Association for Computational Lin-

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2023. Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12506–12521, Singapore. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. ICML'24. JMLR.org.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Yotam Wolf, Noam Wies, Dorin Shteyman, Binyamin Rothberg, Yoav Levine, and Amnon Shashua. 2024. Tradeoffs between alignment and helpfulness in language models with representation engineering. *arXiv preprint arXiv:2401.16332*.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3997–4003. International Joint Conferences on Artificial Intelligence Organization. Main track.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages

Junwei Zhang, Huamin Feng, Biao Liu, and Dongmei Zhao. 2023a. Survey of technology in network security situation awareness. *Sensors*, 23(5).

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023b. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. 2024. Weak-to-strong jailbreaking on large language models. *arXiv preprint arXiv:2401.17256*.

Chujie Zheng, Fan Yin, Hao Zhou, Fandong Meng, Jie Zhou, Kai-Wei Chang, Minlie Huang, and Nanyun Peng. 2024. On prompt-driven safeguarding for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 61593–61613. PMLR.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. On the role of attention heads in large language model safety. In *The Thirteenth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023a. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023b. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.