Inference-Time Control for Trustworthy Large Language Models

Yuyang Bai[†‡1] **Zheyuan Liu**[†‡2] **Han Yan**[‡2] **Zhangchen Xu**[‡3] **Yixin Wan**[‡4]
Canyu Chen[‡5] **Zehong Wang**[‡2] **Xiangchi Yuan**[‡6] **Yue Huang**[‡2] **Guangyao Dou**[‡7]
Yuji Zhang[8] **Hangxiao Zhu**[1] **Zhuofeng Li**[1] **Manling Li**[5] **Xiangliang Zhang**[2]
Mohit Bansal[9] **Sanmi Koyejo**[10] **Kai-Wei Chang**[4] **Yu Zhang**[§1] **Meng Jiang**[§2]

[1]Texas A&M University [2]University of Notre Dame [3]University of Washington
[4]University of California, Los Angeles [5]Northwestern University
[6]Georgia Institute of Technology [7]Johns Hopkins University
[8]University of Illinois at Urbana-Champaign [9]University of North Carolina at Chapel Hill
[10]Stanford University

{ybai, zhuofengli, hangxiao, yuzhang}@tamu.edu,
{zliu29, zwang43, yhuang37, xzhang33, mjiang2}@nd.edu,
yanh8284@gmail.com, zxu9@uw.edu,
{elaine1wan, kwchang}@cs.ucla.edu,
canyuchen@u.northwestern.edu, manling.li@northwestern.edu,
xyuan300@gatech.edu, gdou1@jhu.edu,
yujiz@illinois.edu, mbansal@cs.unc.edu, sanmi@stanford.edu

Abstract

Once a large language model is released, training-time alignment is hard to revise; yet deployment introduces context-specific risks that the original training cannot anticipate: evolving safety policies, jurisdictional constraints, retrieval contamination, and adaptive adversarial prompting. In this paper, we unify inference-time techniques for trustworthy generation across safety, privacy, fairness, and factuality under a single framework: the *inference-time control plane*, with three tiers of intervention—*External Controls* (Context Engineering, Guardrails, Decoding Strategies), which act around the model; *Internal Manipulations* (Representation Engineering, Unlearning, Pruning), which act inside it; and *System-Level Orchestration* (Multi-Agent Systems), which coordinate several models. We also introduce a meta-axis evaluation framework that crosses the four trustworthiness dimensions with five evaluation axes (effectiveness, locality, generality, interpretability, efficiency), and describe representative metrics at each intersection. We identify four cross-cutting open problems: brittleness under adaptive adversaries, the control–utility tradeoff, verification of removal, and the composition of layered interventions. A curated paper list is available at https://github.com/leopoldwhite/Awesome-Inference-Time-Trustworthiness.

[†]Project Leader.
[‡]Major Contributor.
[§]Corresponding Author.

Preprint. Under review.

Contents

1 Introduction

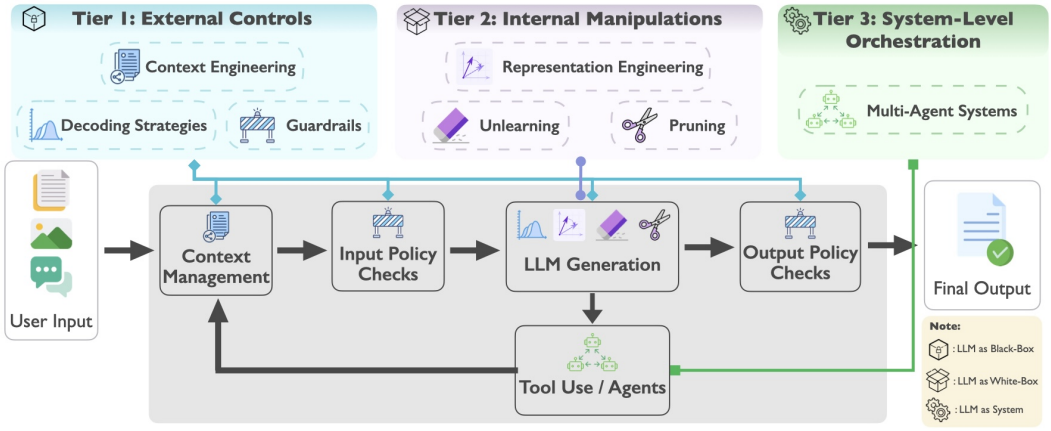

Figure 1: Taxonomy (top) and attachment points in an inference-time pipeline (bottom). We organize seven inference-time methods into three color-coded tiers—*Tier 1: External Controls* (**cyan**, black-box: Context Engineering, Guardrails, Decoding Strategies), *Tier 2: Internal Manipulations* (**lavender**, white-box: Representation Engineering, Unlearning, Pruning), and *Tier 3: System-Level Orchestration* (**green**, system-level: Multi-Agent Systems). Each tier mapped to its point of intervention in the generation pipeline.

Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) have rapidly moved from research prototypes to deployed systems in domains such as healthcare, law, and finance. Models are now accessed primarily through APIs, regulatory frameworks like the EU AI Act are imposing new behavioral requirements, and new vulnerabilities are discovered continuously after deployment. Existing work has largely emphasized train-time trustworthiness, where safety, fairness, privacy, or factuality objectives are encoded during parameter learning through data curation, supervised alignment, or preference optimization. These methods aim to shape the model's default behavior globally and persistently by modifying what the model internalizes.

In this paper, we distinguish this perspective from inference-time trustworthiness, which concerns mechanisms that regulate model behavior at runtime after training has been completed. Rather than changing the model's learned parameters alone, inference-time methods operate on the generation process through contextual control, decoding-time intervention, representation engineering, internal steering, or multi-agent verification. This distinction extends beyond considerations of efficiency or convenience. More fundamentally, train-time and inference-time methods differ in where control is exercised, how persistent the intervention is, and what level of guarantees it can provide. Train-time methods primarily shape a model's behavioral prior, whereas inference-time methods determine how that prior is enacted, constrained, or corrected in a specific interaction context.

These two forms of trustworthiness are complementary. Train-time alignment is well suited for establishing broad and stable default behavior, but deployment introduces context-specific risks that cannot always be fully included during training, including changing policies, user-specific constraints, retrieval contamination, adversarial prompting, and jurisdiction-dependent requirements. Inference-time mechanisms provide a second control plane that is local, adaptive, and often more directly auditable at the level of individual interactions, as evidenced by guardrail families that expose per-query policy identifiers, risk scores, and decision traces [22, 120, 103, 136]. They offer a complementary control layer that can steer, constrain, or monitor model behavior after deployment. Because they are often lightweight and modular, they are well suited to dynamic and context-specific trustworthiness requirements. Conversely, inference-time control alone cannot substitute for robust underlying models, because runtime interventions may be brittle, reactive, or incomplete without strong train-time foundations—for example, adversarial probing can recover supposedly unlearned knowledge [107, 32], and guardrails remain susceptible to adaptive jailbreaks that evade fixed detectors [125]. A full account of trustworthy LLM deployment therefore requires treating trustworthiness as a multi-layer property spanning both training and inference.

The necessity of this multi-layered approach is especially salient in high-stakes deployments such as clinical decision support, legal drafting, and financial advising, where even low-frequency trustworthiness failures can translate into concrete harm and where jurisdiction-specific constraints make a single global default behavior insufficient. As general-purpose reasoning capabilities continue to strengthen, a complementary framing has gained traction in the recent agent literature. Often termed harnessing, this approach treats deployment-time trustworthiness as the problem of wrapping capable but imperfectly aligned models in runtime scaffolding, such as context management, policy enforcement, and agent coordination. This concept is commonly summarized by the equation $\mathrm{Agent} = \mathrm{Model} + \mathrm{Harness}$. This framing maps naturally onto two tiers of our taxonomy: Tier 1 External Controls and Tier 3 System-Level Orchestration. However, it excludes an equally important class of interventions that act on the model's internal computations, such as Representation Engineering, Unlearning, and Pruning. These internal methods become essential when trustworthiness requirements cannot be met by surface scaffolding alone, requiring us instead to bound what the model can recall, represent, or activate in the first place. We therefore adopt a broader inference-time scope that unifies harness-style external controls with internal manipulations under a single framework, capturing the full stack of deployment-time intervention points available after training concludes.

Many inference-time techniques were first developed to improve model performance, such as reducing latency, improving fluency, or grounding outputs through retrieval. More recently, they have also been adapted to support trustworthiness during deployment, including safety, privacy, fairness, and factuality. This shift calls for a more unified view of the field. In this paper, we organize inference-time trustworthiness methods by the mechanism and locus of control they introduce during generation. As shown in Figure 1, we group them into three tiers: **External Controls**, **Internal Manipulations**, and **System-Level Orchestration**. This perspective highlights shared design patterns and clarifies trade-offs in granularity, invasiveness, and modularity. These tiers differ in where and how they intervene, ranging from black-box control around the model, to white-box intervention inside the model, to coordinated control across multiple agents.

Tier 1: External Controls. Black-box methods that intervene at the *context assembly*, *input/output policy checks*, and *decoding* stages of the pipeline:

- **Context Engineering** (Section 2.1): Strategic prompt design through rules, instructions, or few-shot exemplars to guide outputs without modifying model parameters.
- **Guardrails** (Section 2.2): External modules that inspect inputs or outputs against safety or policy constraints, blocking, redacting, or regenerating content when violations occur.
- **Decoding Strategies** (Section 2.3): Manipulation of token-level distributions during generation to promote desired attributes or suppress undesired ones.

Tier 2: Internal Manipulations. White-box methods that operate within the *LLM generation* stage of the pipeline:

- **Representation Engineering** (Section 3.1): Direct modification of internal activations by adding or subtracting steering vectors associated with specific concepts.
- **Unlearning** (Section 3.2): Targeted removal of information, behaviors, or biases from a pre-trained model to fulfill data-forgetting requirements or disable harmful capabilities.
- **Pruning** (Section 3.3.1): Post-training removal of weights, neurons, or attention heads, originally for efficiency but now increasingly explored for trust-related effects.

Tier 3: System-Level Orchestration. A single category, **Multi-Agent Systems** (Section 4.1), in which a *tool use / agents* loop spans the entire pipeline through coordinated agent interactions such as debate, cross-verification, and role specialization.

These three tiers form a defense-in-depth view of inference-time trustworthiness. External controls provide lightweight and modular policy enforcement. Internal manipulations enable more direct and fine-grained behavioral intervention. System-level orchestration introduces an additional layer of reliability through collaborative reasoning and feedback. Across this taxonomy, we focus on four dimensions of trustworthiness: **Safety**, which concerns preventing harmful, biased, or malicious outputs and defending against misuse such as jailbreaks; **Privacy**, which concerns limiting training-data

Table 1: Comparison with related work on inference-time trustworthiness metrics and methods.

Prior Work	Inference-time Metrics				Inference-time Methods						
	Trustworthiness Dimensions				External Controls			Internal Manipulations			System-Level Orchestration
	Safety	Privacy	Fairness	Factuality	Context Engineering	Guardrails	Decoding	Representation Engineering	Unlearning	Pruning	Multi-Agent Systems
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Donisch et al. [31]	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓	✗
Kumar et al. [81]	✓	✓	✗	✓	✗	✓	✓	✓	✗	✓	✓
Tie et al. [142]	✓	✗	✓	✓	✗	✗	✓	✓	✗	✓	✓
Huang et al. [63]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Liu et al. [102]	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗
Wan et al. [151]	✓	✗	✓	✓	✓	✗	✗	✓	✗	✗	✗
Yu et al. [170]	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
Barez et al. [6]	✓	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗
Liu et al. [105]	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✗

leakage and protecting user inputs during inference; **Fairness**, which concerns reducing systematic disadvantages toward individuals or groups; and **Factuality**, which concerns grounding model outputs in verifiable knowledge and reducing hallucinations. These three tiers and four dimensions together define the inference-time control plane, and we analyze how its components trade off and compose under deployment constraints.

1.1 Related Work

This work sits at the intersection of three research strands. **Efficiency-focused work** on inference-time optimizations such as quantization, distillation, speculative decoding, and Mixture-of-Experts deployment [31, 99, 196] targets computational performance rather than trustworthy behavior. **Post-training capability research** on prompting and reinforcement learning aims to make models more capable reasoners or tool users [142, 81]; deployment-time enforcement of safety, privacy, or fairness is peripheral. **Trustworthy AI research** either takes a broad view [63, 102] or focuses on specific facets such as text-to-image bias [151], autonomous agent safety [170], and machine unlearning [105, 6]. To our knowledge, no prior work provides a unified framework that treats *inference-time methods as a control plane* for trustworthiness across safety, privacy, fairness, and factuality.

1.2 Contributions and Organization

Our key contributions are:

- **A Pipeline-Grounded, Three-Tier Framework.** We propose a three-tier framework grounded in the generation pipeline (Figure 1). The framework covers seven inference-time control methods and groups them by their intervention point, access requirements, and composability within a defense-in-depth design.
- **Cross-Dimensional Analysis.** We analyze how each category addresses Safety, Privacy, Fairness, and Factuality, clarifying capabilities, limitations, and trade-offs under deployment constraints.
- **Cross-Cutting Open Problems.** We identify open problems on scalability, evaluation, and composition that cut across the framework, and discuss directions for future work.

The remainder is organized as follows. **Sections 2–4** cover the three tiers in turn. **Section 5** analyzes evaluation benchmarks and trade-offs. **Section 6** discusses open problems and limitations, and **Section 7** concludes.

2 External Controls

External controls treat the model as a black box, shaping its behavior by manipulating inputs, the decoding process, or outputs—without accessing or modifying internal weights or activations. These methods are the most modular and widely applicable, as they require no white-box access and can be deployed on proprietary, API-only models. In the inference-time pipeline (Figure 1), they attach to the context assembly, input policy checks, decoding, and output policy checks stages.

2.1 Context Engineering

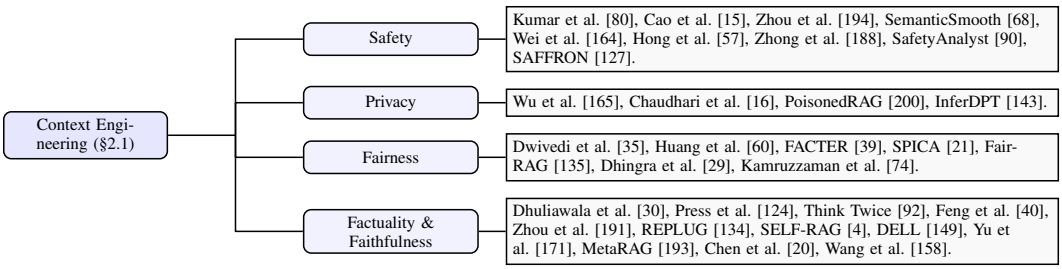

Figure 2: A taxonomy of Context Engineering for enhancing LLM trustworthiness.

Within the landscape of inference-time methods for trustworthy LLMs, context engineering (CE) has emerged as a central paradigm. Instead of modifying model parameters, CE emphasizes the deliberate design and orchestration of the information presented to the model at inference time. This approach is particularly important for trustworthiness: by shaping the context, one can systematically influence factuality, safety, fairness, etc, without modifying model's parameters.

2.1.1 Core Principle

Context engineering is the systematic discipline of designing, optimizing, and managing the information payload provided to LLMs at inference time. Unlike early prompt engineering, which treated the context as a flat string, CE models the context as a structured composition of four components: directive instructions encompassing rules and constraints; knowledge inputs such as retrieved documents or prior dialogue; interaction mechanisms including tools, APIs, or system affordances; and the live user query specifying the task at hand. An assembly function integrates these heterogeneous elements into the final sequence consumed by the model.

The CE framework rests on three foundational pillars that together form a pipeline. Context retrieval and generation is responsible for sourcing task-relevant material through prompt design, retrieval from external knowledge bases, and adaptive assembly strategies. Context processing transforms raw material into model-ready input through compression, refinement, and multimodal or structured integration. Context management ensures coherence and efficiency across stages by dynamically organizing, prioritizing, and updating contextual elements under system constraints such as the maximum context length. Together, these pillars support higher-level methods such as Retrieval-Augmented Generation (RAG), persistent memory, tool-augmented reasoning, and multi-agent collaboration.

2.1.2 Applications

Safety Context engineering promotes safety by controlling the informational boundary within which a model operates, encoding explicit constraints at the prompt level, grounding generation in trusted external sources, and embedding verification routines into the reasoning process. Prompt engineering offers a lightweight first layer of defense, ranging from certifiable defenses with robustness guarantees against adversarial prompts [80] and robust alignment prompting [15] to baseline defenses such as paraphrasing, in-context refusals, and perplexity checks [194], as well as smoothing-based methods like SemanticSmooth that guard against both token- and prompt-level attacks [68]. Retrieval-augmented generation further enhances safety by grounding responses in verifiable external knowledge, reducing hallucinations and lowering the risk of unsafe misinformation [164, 57], though vulnerabilities such as corpus poisoning attacks that insert adversarial passages into knowledge bases remain a concern [188]. Reasoning-oriented approaches complement these strategies by enabling models to regulate their own outputs during inference through structured step-by-step verification [90] and safety-aware scoring that resists jailbreaks at scale [127].

Privacy Context engineering for privacy focuses on controlling what enters the model's context. Instead of modifying parameters, privacy is preserved by assembling the context such that sensitive elements are masked, filtered, or abstracted while retaining task-relevant information. Recent work shows that CE can both introduce and mitigate privacy risks. Prompts themselves may leak sensitive

training data through property and membership inference attacks [165], and RAG pipelines are vulnerable to poisoned documents that can exfiltrate private passages or bias model outputs [16]. On the defense side, certified privacy-preserving mechanisms for RAG offer robustness against adversarial manipulation [200], while frameworks for selective retrieval and privacy-aware generation control personal data exposure [143]. Together, these studies underscore that privacy risks often stem from how external context is engineered into models, and effective mitigation requires careful control, sanitization, and certification of injected information.

Fairness Bias in LLMs often arises from skewed training data and emergent model behavior, and context engineering provides a practical means of mitigation without retraining. Carefully designed prompts have been shown to reduce bias, from employing in-context learning to mitigate gender stereotypes [35] to frameworks for testing and reducing social bias in code generation through iterative, feedback-driven prompting [60] and embedding explicit fairness constraints in prompts [39]. Retrieval-augmented methods enhance fairness by grounding outputs in curated, balanced evidence, with fairness-aware retrieval mechanisms prioritizing diverse and representative sources to reduce stereotype reinforcement [21, 135]. A third line of work integrates structured reasoning through chain-of-thought [29] or deliberative reasoning [74] to enforce fairness constraints, decomposing decisions into interpretable steps that enable bias detection and correction during inference.

Factuality A core challenge in trustworthy AI is mitigating factual errors and hallucinations, and context engineering offers several complementary strategies. Prompt engineering improves factual consistency through structured instructions, including self-detection prompts that require models to evaluate multiple candidate answers before finalizing a response [30, 124, 92], abstention-oriented prompting that encourages models to decline uncertain answers in multilingual settings [40], and opinion-based or counterfactual demonstrations that reduce reliance on parametric knowledge [191]. Retrieval-augmented generation enhances factual grounding by supplementing LLMs with external documents, with frameworks enabling adaptive and reflective retrieval such as SELF-RAG [4], integration of retrieval with proxy explanations for misinformation detection [149, 171, 134], and metacognitive loops that detect insufficient or conflicting knowledge and adjust retrieval accordingly [193]. Memory-based methods further improve factuality by allowing models to retain and update knowledge across interactions through external memory buffers where user feedback persists across sessions [20] and modular memory systems combining episodic traces and semantic caches [158].

2.2 Guardrails

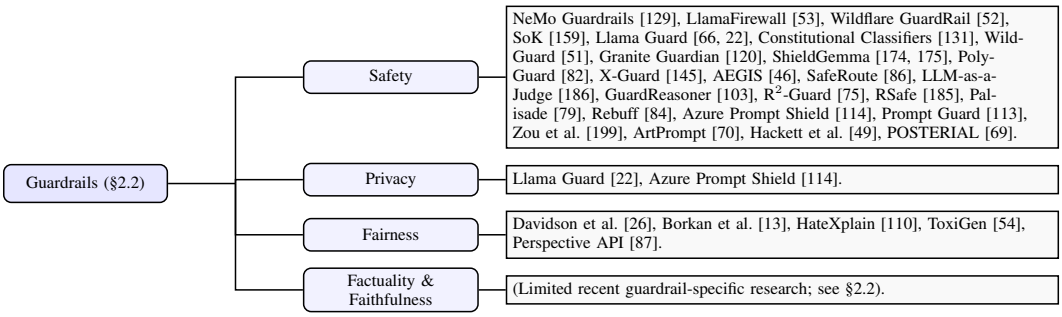

Figure 3: A taxonomy of Guardrails for enhancing LLM trustworthiness.

2.2.1 Core Principle: Modular and External Control

Guardrails provide a modular and external mechanism for controlling LLM behavior, operating as independent components that monitor and filter inputs or outputs without modifying the model's parameters. Unlike training-time interventions, guardrails function post-deployment, treating the LLM as a black-box system. This external positioning enables rapid deployment and updates, allowing systems to quickly adapt to evolving safety requirements and emerging threats. Their modular design supports flexible integration into existing pipelines, where multiple guardrails can be layered or combined to target specific risks. As external controls, guardrails ensure trustworthy outputs by

intercepting potentially harmful content before it reaches the user or by regenerating outputs when violations are detected. While this principle draws conceptually from traditional content moderation systems, it is now tailored to the dynamic and interactive nature of LLMs, emphasizing flexibility and minimal interference with the model's core functionality.

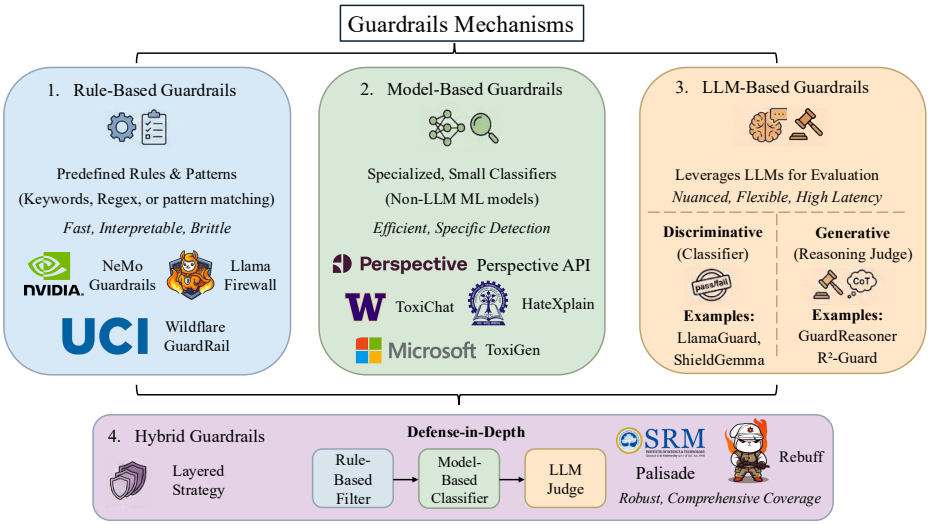

Figure 4: A typology of guardrail mechanisms for LLM trustworthiness. The four categories—Rule-Based, Model-Based, LLM-Based (discriminative and generative), and Hybrid—are distinguished by their underlying detection approach, with representative systems shown for each. Hybrid guardrails combine multiple mechanism types in a layered "defense-in-depth" strategy for comprehensive coverage.

2.2.2 A Typology of Guardrail Mechanisms

Guardrail mechanisms can be classified according to their underlying approach, spanning from simple deterministic techniques to advanced LLM-driven systems.

Rule-Based Guardrails Rule-based guardrails rely on predefined rules, such as keyword lists, regular expressions, or pattern matching, to detect and filter prohibited content. For example, NeMo Guardrails [129] introduces programmable rails for dialog flow, topic blocking, and tool use; LlamaFirewall [53] incorporates customizable scanners including regex-based code filters for agentic workflows; and Wildflare GuardRail [52] integrates modular rule-based wrappers as part of a multi-stage pipeline. While transparent and low-latency, these systems are brittle against novel adversarial strategies specific to LLMs, such as prompt injection and jailbreak attacks [159].

Model-Based Guardrails Model-based guardrails employ smaller, specialized classifiers—typically non-LLM models such as neural networks or fine-tuned smaller transformers—for nuanced detection of safety violations. Early work focused on hate speech and toxicity detection, addressing challenges such as distinguishing hate speech from offensive content [26], fairness-oriented evaluation for toxicity detectors [13], and adversarially generated datasets for implicit toxicity [54]. These insights informed practical deployments such as the Perspective API [87], which detects toxicity across languages, and explainable models like HateXplain [110] that provide human rationales alongside labels.

LLM-Based Guardrails LLM-based guardrails employ another large language model to evaluate and filter generated content, capturing nuance, abstraction, and intent that smaller classifiers often miss. Discriminative guardrails use LLMs as safety classifiers: Llama Guard [66, 22] detects unsafe multimodal content categories, while Constitutional Classifiers [131], WildGuard [51], Granite Guardian [120], and ShieldGemma [174, 175] provide fine-tuned drop-in safety classifiers. Recent work has expanded to multilingual coverage with PolyGuard [82] and X-Guard [145], while ensem-

ble approaches like AEGIS [46] and adaptive routing via SafeRoute [86] reflect a shift toward scalable, real-time solutions. Generative (reasoning) guardrails employ an LLM as a deliberative judge that critiques and justifies another model's output. Building on the LLM-as-a-judge paradigm [186], recent methods incorporate explicit reasoning chains before producing verdicts, such as GuardReasoner [103], R^2-Guard [75] with knowledge-enhanced logical reasoning, GuardAdvisor [65] with a soft-gating pipeline, and RSafe [185] with adaptive guided safety reasoning.

Hybrid Guardrails Hybrid guardrails combine multiple mechanisms for comprehensive coverage through a "defense-in-depth" strategy. For example, Palisade [79] proposes a three-layer pipeline combining a rule-based filter, an ML classifier, and a companion-LLM check, while Rebuff [84] integrates heuristics, canary tokens, a database of known attacks, and an LLM-based check.

2.2.3 Applications

Safety Guardrails are predominantly applied in safety, where they serve as critical defenses against emerging threats inherent to LLMs. On the input side, guardrails detect and block attacks such as jailbreaks that aim to bypass alignment constraints [199, 70] and prompt injections where adversaries embed malicious instructions to hijack model outputs [49, 69]. Systems like Azure Prompt Shield [114] and Prompt Guard [113] employ adversarial-input detectors to identify injection attempts and block unsafe prompts. On the output side, guardrails evaluate and filter model generations before delivery to end users, with systems like Llama Guard [22] enforcing a taxonomy of unsafe categories—including violent crimes, sex crimes and child exploitation, defamation, hate speech, self-harm, and election misinformation—by classifying outputs and blocking or regenerating unsafe completions.

Privacy In the privacy domain, guardrails protect against the disclosure of personal or sensitive information. Output-side guardrails can classify and block generations that contain personally identifiable information, confidential data, or intellectual property violations. Systems such as Llama Guard [22] include disclosure of sensitive information as an explicit unsafe category, while Azure Prompt Shield [114] detects attempts to exfiltrate private data through prompt injection. However, guardrail-specific research for privacy remains limited compared to safety, and most privacy protections are currently bundled within broader safety taxonomies rather than addressed by dedicated guardrail mechanisms.

Fairness Guardrails for fairness have historical roots in pre-LLM content moderation, where classifiers were designed to detect and reduce bias in generated content [13, 110]. Toxicity and hate speech detectors such as Perspective API [87] and ToxiGen classifiers [54] serve as fairness-oriented guardrails by flagging content that disparages or stereotypes specific demographic groups. In the LLM era, however, fairness-specific guardrail research has received less attention, likely due to the assumption that post-training alignment methods such as RLHF improve baseline fairness, leaving a gap for dedicated fairness enforcement mechanisms at inference time.

Factuality Guardrails for factuality represent the least developed dimension in current research. While output-side guardrails could in principle verify factual claims against trusted knowledge bases or flag low-confidence generations, there is limited recent work on dedicated factuality-focused guardrail mechanisms for LLMs. This gap presents an opportunity for future research to develop fact-checking guardrails that complement retrieval-augmented and decoding-based approaches to factual grounding.

2.3 Decoding

2.3.1 Core Principle: Real-Time Logit and Probability Manipulation

Decoding strategies involve real-time manipulation of the token probability distribution or logits at each generation step to steer the model's outputs toward desired characteristics. At each step, the model produces a probability distribution over its vocabulary conditioned on the preceding tokens, and decoding interventions modify this distribution—by suppressing, amplifying, or re-weighting specific token probabilities—before sampling the next token. This approach allows for fine-grained

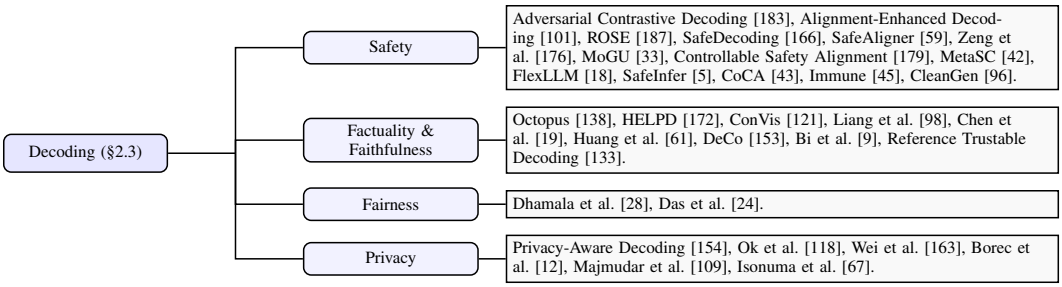

Figure 5: A taxonomy of Decoding Strategies for enhancing LLM trustworthiness.

control over output without modifying the model's parameters, making it computationally efficient and adaptable to diverse trustworthiness requirements.

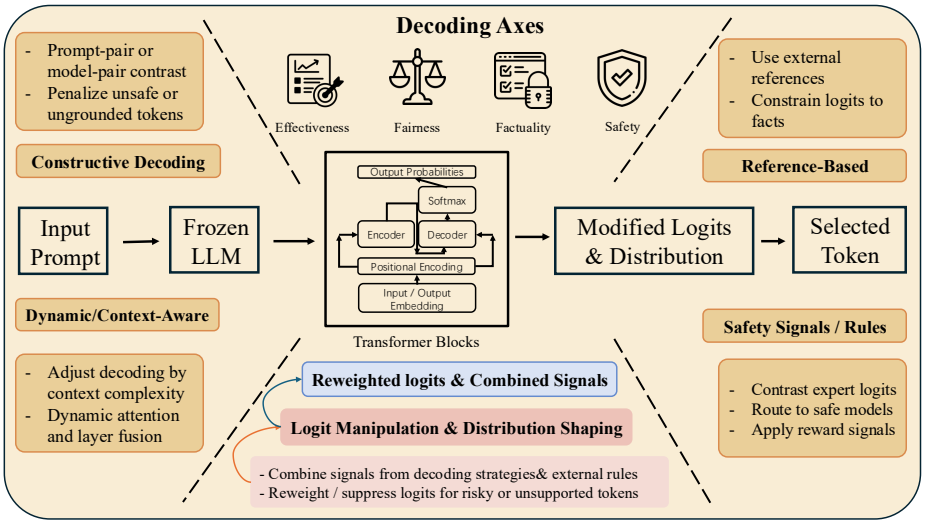

Figure 6: Overview of trustworthiness-oriented decoding strategies. A frozen LLM produces logits that are modified at each autoregressive step by three complementary families of interventions before token sampling. **Contrastive Decoding** (balance icon) contrasts outputs from paired prompts or paired models—e.g., safe versus unsafe prompts, or an expert against an amateur model—to penalize unsafe or ungrounded tokens. **Dynamic/Context-Aware Decoding** (neural-network icon) adjusts decoding hyperparameters, attention weights, or layer fusion on the fly, based on context complexity or intermediate signals. **Reference-Based Decoding** (document-with-shield icon) constrains logits toward facts or safety policies sourced from external references such as retrieved documents or rule sets. These three families feed into a shared *Logit Manipulation & Distribution Shaping* stage, optionally combined with external knowledge or safety signals, which reweights or suppresses risky, unsupported, or unsafe tokens before the next token is sampled.

2.3.2 Applications

Safety Decoding strategies for safety center on inference-time intervention to prevent harmful outputs without altering model weights, and are crucial for defending against adversarial attacks such as jailbreaks. Contrastive decoding enhances safety by comparing logits from safe and unsafe distributions: methods like Adversarial Contrastive Decoding [183] and Alignment-Enhanced Decoding [101] use opposite or malicious prompts to penalize unsafe continuations, while SafeDecoding [166] and SafeAligner [59] contrast logits from safety-trained expert models against base or unsafe models, and ROSE [187] applies reverse prompt contrastive decoding with step-by-step correction [176]. External guidance approaches use separate models or signals to steer generation, such as MoGU's dynamic routing between a usable and a safe LLM [33], inference-time adaptation

through safety configurations [179], and meta-critique optimization of safety specifications [42]. Context-adaptive and dynamic defenses adjust safety measures based on specific inputs, including moving target defense through dynamic decoding hyper-parameters [18] and context-adaptive safety amplification [5]. Specialized defense work addresses safety in multimodal LLMs through output distribution calibration [43] and safety reward model alignment [45], as well as countering backdoor attacks via speculative decoding to identify suspicious tokens [96].

Factuality Diverse decoding strategies have been developed to mitigate hallucinations and ensure context-aware outputs in LLMs and MLLMs. Contrastive decoding reduces hallucinations by contrasting output distributions derived from original and perturbed inputs: Octopus [138] adapts to input complexity with noise, HELPD [172] employs hierarchical feedback with vision-enhanced penalty decoding, ConVis [121] penalizes hallucinated content via text-to-image reconstruction, and further methods re-balance distributions to prioritize visual grounding [98] or integrate multi-layer fusion with contrastive decoding [19]. Context-aware decoding dynamically adjusts strategies to ensure faithfulness to input context, such as dynamically adjusting attention mechanisms to prioritize context-relevant tokens [61] and adaptively selecting preceding layers to correct output logits [153]. Reference-based decoding enhances factual consistency by contrasting knowledge states to improve confidence on edited facts [9] or using reference information to ensure output consistency without additional training [133].

Privacy Decoding strategies address privacy concerns by targeting two objectives: reducing privacy leakage and reducing memorization. For leakage prevention, Privacy-Aware Decoding [154] filters outputs in RAG systems to prevent disclosure of sensitive retrieved data, selective teacher supervision during decoding avoids leaking data from untrusted sources [118], and privacy vulnerabilities in speculative decoding have been identified where accelerated inference may increase leakage risks [163]. For memorization reduction, nucleus sampling has been shown to only modestly reduce memorization due to peaked output distributions [12], differential privacy applied during decoding by adding noise to logits provides formal guarantees [109], and contrastive generation encourages novel outputs to reduce reliance on memorized training data [67].

Fairness While decoding strategies have been explored for safety, factuality, and privacy, their impact on fairness remains a nascent and under-explored domain. Dhamala et al. [28] provide a pioneering analysis of how decoding parameters such as top-k, top-p, and temperature sampling influence fairness in open-ended text generation, demonstrating that specific hyperparameter configurations can exacerbate or mitigate group-level disparities. Similarly, Das et al. [24] explore bias across the entire decoder hyperparameter space, revealing critical trade-offs between bias reduction and generation quality. These studies highlight fairness as an emerging frontier in decoding methods, where systematic methodologies and robust interventions remain a pressing need.

3 Internal Manipulations

Internal manipulations require white-box access to the model. They intervene directly in the model's computation, for example by modifying activations during a forward pass, selectively removing targeted behaviors or knowledge in context, or pruning architectural components. Compared with external controls, these methods often provide finer-grained and more persistent behavioral changes. In return, they require white-box access to internal representations or model components during generation. In the inference-time pipeline shown in Figure 1, they operate at the LLM generation stage by acting on internal representations or model components.

3.1 Representation Engineering

3.1.1 Core Principle

Representation Engineering (RepE) is an inference-time control paradigm for LLMs that directly modifies internal activations, often in the residual stream at selected layers, to steer high-level behaviors such as refusal [184, 44] or truthfulness [91]. Unlike methods that focus on individual neurons or fully specified circuits, RepE treats concepts as directions or low-rank subspaces in population activations and manipulates them directly. This top-down view supports both probing and

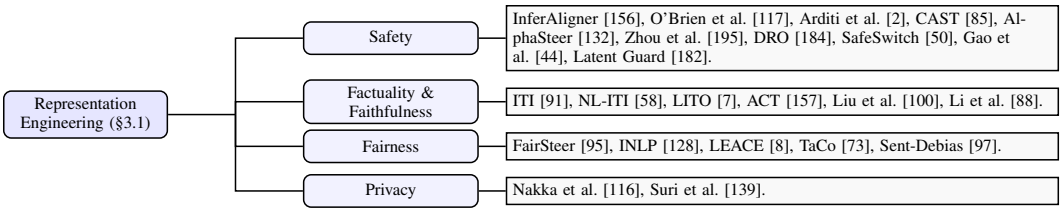

Figure 7: A taxonomy of Representation Engineering for enhancing LLM trustworthiness.

steering of abstract properties within a single forward pass [198]. In practice, RepE identifies how a concept is encoded in activations, represents it as a direction or subspace, and applies targeted control through activation steering or related interventions. This makes RepE a flexible, efficient, and interpretable way to shape model behavior while largely preserving overall capabilities [161].

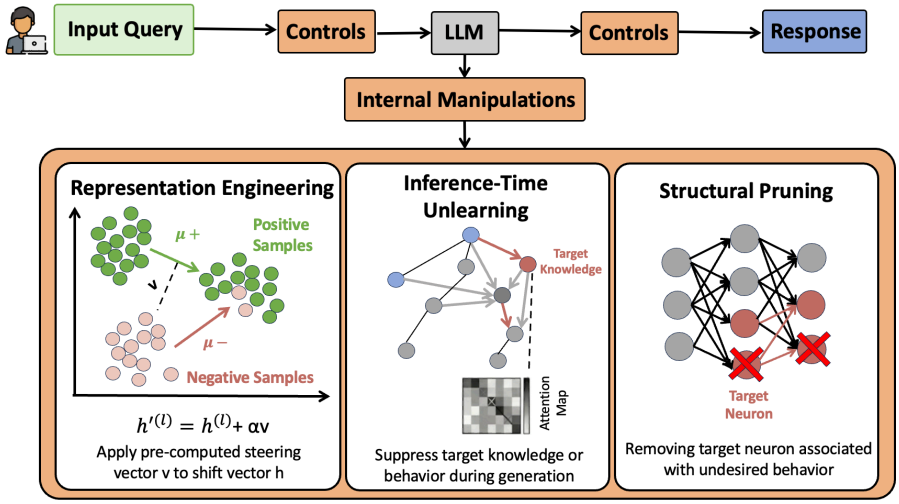

Figure 8: Overview of internal manipulation methods at inference time, including representation engineering, inference-time unlearning, and structure-level interventions such as pruning.

3.1.2 Mechanism

RepE typically works by constructing steering vectors, which are directions in activation space associated with a target concept, and injecting them during inference. This process has two stages: identifying concept-relevant representations and applying them at selected layers.

Let $h^{(\ell)} \in \mathbb{R}^d$ denote the hidden representation at layer ℓ in the residual stream, where d is the representation dimension. Given two prompt sets, X^+ for the target concept and X^- for its contrast, we compute their mean activations at layer ℓ:

$$\mu^+ = \frac{1}{|X^+|} \sum_{x \in X^+} h^{(\ell)}(x), \qquad \mu^- = \frac{1}{|X^-|} \sum_{x \in X^-} h^{(\ell)}(x).$$

The steering vector is then defined as

$$v = \mu^+ - \mu^-,$$

representing the concept as a direction in activation space. More generally, dimensionality reduction methods such as PCA [72], SVD [47], and SAE, as well as optimization-based methods, can be used to extract low-rank subspaces that capture the concept beyond a single direction [71, 144, 189].

During inference, the model is steered by modifying the residual activation before it is passed to later layers. For an input x with hidden state $h^{(\ell)}(x)$, the steered representation is

$$\tilde{h}^{(\ell)}(x) = h^{(\ell)}(x) + \alpha v,$$

where $\alpha \in \mathbb{R}$ controls the strength and direction of the intervention. Positive α strengthens the target concept, while negative α suppresses it. This intervention is lightweight because it adds no trainable parameters and only requires a vector addition at inference time. By adjusting α, model behavior can be continuously controlled without retraining or weight updates [144]. Some extensions replace the additive form with affine transformations,

$$\tilde{h}^{(\ell)}(x) = Wh^{(\ell)}(x) + b,$$

where (W, b) are low-rank steering operators that provide more expressive but still efficient control [71, 189].

3.1.3 Applications

Safety Representation engineering has become an important tool for safety control at inference time. Existing work mainly uses activation-level steering to modulate refusals, suppress harmful outputs, or reduce false refusals. Some methods inject safety vectors during generation, either through cross-model guidance or harm-specific steering, to improve safety while preserving utility [156]. Others use sparse autoencoders to identify refusal-related features or derive steering directions for detoxification, which improves controllability but may introduce utility trade-offs [117]. Mechanistic studies further suggest that refusal behavior can often be traced to a small set of latent directions or attention components, which enables more targeted interventions such as conditional activation steering and false-refusal ablation [2, 85, 132, 195].

This line of work also provides a clearer view of how safety steering interacts with jailbreak robustness. Prior studies show that generic safety prompts often over-shift representations toward refusal, which can increase blanket refusals even when the model can distinguish harmful from benign inputs [184]. In response, later methods study more selective steering strategies and more robust safety control under adversarial prompting [50, 44, 182]. Overall, these results show that representation engineering offers a lightweight and increasingly precise toolkit for inference-time safety control.

Fairness Fairness can also be improved through interventions on internal representations rather than full retraining. In inference-time settings, activation steering uses bias-related directions to modify hidden states and reduce demographic or ideological skew. Recent methods further adapt the intervention layer and strength to improve fairness with limited utility loss [95]. Related work studies concept erasure, which removes protected-attribute information from embeddings or activations through projection or transformation. Representative methods include INLP, LEACE, TaCo, and Sent-Debias [128, 8, 73, 97]. Together, these studies suggest that representation-level interventions can help reduce bias without full model retraining.

Privacy Representation engineering also plays a dual role in privacy, as activation steering can both expose and mitigate leakage. On the attack side, Nakka et al. [116] identify refusal-related attention heads for sensitive attributes and steer a small subset of them at inference time to bypass safeguards, leading to high disclosure rates across several LLMs. On the defense side, Suri et al. [139] show that activation steering can suppress verbatim memorization on a controlled benchmark with limited quality loss. Together, these results suggest that inference-time steering can serve both privacy auditing and mitigation without retraining.

Factuality and Faithfulness A major line of work in representation engineering steers internal activations at inference time to improve truthfulness without updating model weights. ITI shifts a small set of attention-head activations along truth-related directions and substantially improves TruthfulQA accuracy on Alpaca [91]. NL-ITI extends this idea with multi-token non-linear interventions and reports further gains over ITI [58]. To avoid fixed intervention strengths, LITO learns an instance-specific schedule and can back off through uncertainty-aware refusal, improving truthfulness while preserving task accuracy [7, 7]. ACT learns a bank of steering vectors and adapts intervention strength by hallucination category, showing gains across the LLaMA family [157]. Other studies provide a more mechanistic view by identifying global truth-related structure in representation space or finer-grained truth neurons associated with truthful output [100, 88].

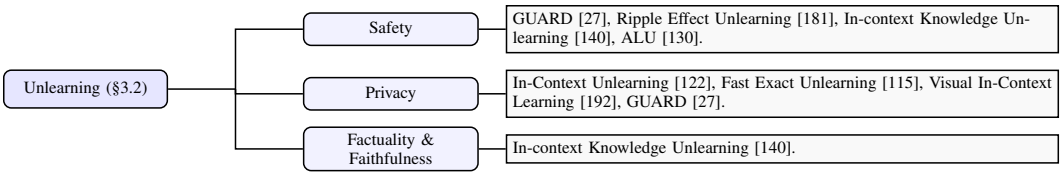

Figure 9: A taxonomy of Unlearning for enhancing LLM trustworthiness.

3.2 Unlearning

3.2.1 Core Principle

LLMs are trained on massive corpora, so controlling what knowledge they retain and express has become increasingly important. At the same time, harmful, private, copyrighted, or otherwise undesirable content may be absorbed during training, raising legal, ethical, and safety concerns. These concerns have made unlearning an important problem for trustworthy LLM deployment. At a high level, unlearning aims to prevent targeted knowledge from influencing model behavior. In inference-time settings, this goal is pursued not by rewriting model parameters, but by intervening during generation to suppress, bypass, or reduce the influence of unwanted knowledge and behaviors. Such methods act as runtime control mechanisms. They limit the model's tendency to express targeted content while preserving its broader capabilities.

3.2.2 Mechanisms

Inference-time unlearning methods can be grouped into two main families: *gradient-based methods* and *influence-based methods*. Gradient-based methods apply dynamic penalties or control signals during the forward pass to steer generation away from targeted knowledge. Influence-based methods instead estimate and remove the contribution of specific in-context examples at inference time.

Gradient-Based Methods Gradient-based unlearning methods suppress targeted knowledge by applying gradient-inspired penalties during the forward pass. Rather than updating model weights, they modify the model's runtime computation to steer generation away from undesirable outputs.

A representative example is **GUARD** [27], which introduces adaptive restriction and detection during generation. By identifying knowledge associated with sensitive concepts and applying penalties in real time, the model is guided away from unsafe continuations while maintaining fluency. Related work also suggests that inference-time forgetting can emerge through runtime control signals that redirect generation toward refusal rather than factual recall [181]. Overall, these methods make unlearning flexible and reversible at deployment time, but they also highlight a central limitation: the targeted knowledge is often suppressed in output rather than fully erased from the model.

Influence-Based Methods Influence-based methods treat forgetting as the removal or reduction of specific contextual influence at inference time. Rather than steering the model globally, they estimate how particular training or in-context examples contribute to a prediction and then cancel, exclude, or replace that effect during generation. In this sense, they frame unlearning as a counterfactual reasoning problem rather than direct intervention on the forward trajectory.

For example, Pawelczyk et al. [122] introduce In-Context Unlearning, where LLMs are taught to disregard selected in-context examples by relabeling or suppressing their contribution, thereby simulating forgetting within the prompt. Building on this idea, Muresanu et al. [115] propose Fast Exact Unlearning, which efficiently removes the influence of in-context training examples from model predictions without weight updates. Although originally motivated by privacy compliance, these methods show how influence estimation can support inference-time unlearning guarantees. This perspective also extends to multimodal settings. Zhou et al. [192] study unlearning in large vision-language models through visual in-context learning and show that selectively excluding visual exemplars at inference time can reduce leakage of sensitive information. Together, these works show that influence-based unlearning is a clean and flexible inference-time mechanism across both text and multimodal generation.

3.2.3 Applications

Although inference-time unlearning methods differ in mechanism, they share a common goal of reducing the influence of harmful, private, or outdated knowledge during generation. Below, we highlight their applications to privacy, safety, and factuality/faithfulness.

Safety A key application of unlearning is the suppression of dangerous or toxic capabilities. Because LLMs are trained on large web corpora, they may retain knowledge of illegal activities or unsafe behaviors that can be elicited at deployment time. Gradient-based methods such as GUARD [27] address this problem by applying adaptive restriction and detection during generation, steering the model away from unsafe continuations without retraining. In-context knowledge unlearning [140] uses unlearning tokens to trigger selective refusal in context, enabling dynamic suppression of unsafe or irrelevant knowledge. Ripple Effect Unlearning [181] further shows that suppressing harmful capabilities, such as bomb-making instructions, can reduce jailbreak success but may also propagate to related benign domains. Multi-agent inference frameworks such as ALU [130] extend this line by formulating safety control through coordinated unlearning-oriented reasoning under diverse jailbreak settings.

Privacy Privacy is one of the clearest motivations for unlearning, especially under regulations such as the right to be forgotten [147, 10]. Because LLMs can memorize sensitive data and disclose it during interaction, inference-time unlearning provides a lightweight way to reduce such leakage without retraining. In-Context Unlearning [122] suppresses the influence of forgotten examples within the prompt, while Fast Exact Unlearning [115] removes the effect of sensitive in-context examples at inference time without changing model parameters. In multimodal settings, Visual In-Context Learning [192] further shows that excluding visual exemplars at inference time can reduce leakage in large vision-language models.

Factuality and Faithfulness Inference-time unlearning also relates to factuality and faithfulness, especially when models rely on outdated, incorrect, or untrusted knowledge during generation. Instead of recalling such content, unlearning can redirect the model toward abstention or explicit forgetting. In-context knowledge unlearning [140] provides a direct example, showing that unlearning tokens can suppress forgotten content and induce explicit "forgotten" responses rather than unsupported answers. This suggests that inference-time unlearning can improve factual reliability by preventing the model from expressing knowledge that should no longer be trusted. At the same time, this connection remains underexplored, and it is still unclear when runtime suppression should be viewed as genuine forgetting rather than controlled abstention. Recent diagnostic frameworks argue that single-value unlearning metrics obscure exactly this distinction, and propose cognitive-diagnosis-style evaluations that jointly probe retention, removal, and transfer effects across difficulty levels [83].

3.3 Pruning

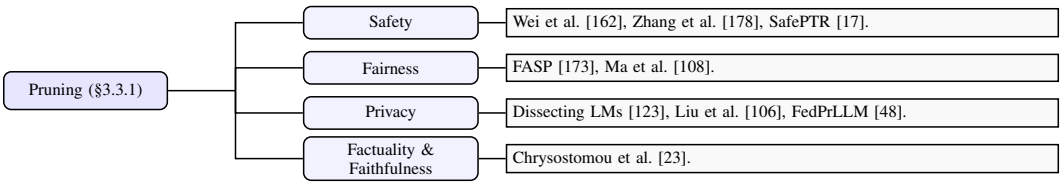

Figure 10: A taxonomy of Pruning for enhancing LLM trustworthiness.

3.3.1 Core Principle

Pruning has traditionally been studied as a tool for improving efficiency by reducing inference cost, memory footprint, and latency. As foundation models move into safety- and regulation-sensitive settings, this framing is no longer sufficient. The key question is no longer only how to make models smaller or faster, but also how to make them safer, fairer, more privacy-preserving, and more reliable.

From this perspective, pruning becomes a form of structural control over model behavior rather than a purely efficiency-driven reduction technique. By selectively suppressing weights, neurons, heads, or subnetworks associated with bias amplification, unsafe generation, or privacy leakage, pruning can reshape the pathways through which the model produces outputs. Methods such as fairness-aware sparsification, safety-constrained pruning, and privacy-oriented structural reduction reflect this shift by embedding trust-related objectives into the model's active computation rather than treating them only as post hoc corrections. Under this view, pruning serves not only to improve efficiency, but also to support leaner and more trustworthy deployment.

3.3.2 Applications

Empirical evidence shows that utility metrics alone can miss important trustworthiness shifts after pruning. Compression may preserve perplexity while still degrading refusal behavior, toxicity control, or other safety properties. This means safety should be treated as a first-class constraint in pruning rather than checked only after compression.

Safety Naive pruning can weaken safety even when utility loss is small. Wei et al. [162] show that removing only $\sim$3% of parameters, or $\sim$2.5% of low-rank components, in safety-critical regions can sharply degrade refusal behavior, suggesting that aligned safety depends on structurally sparse components. Zhang et al. [178] further show that efficiency-oriented activation approximation can also introduce safety vulnerabilities, reinforcing the need to account for safety during compression rather than after it. Recent work therefore explores more targeted pruning strategies. In multimodal LLMs, SafePTR [17] selectively prunes harmful tokens at vulnerable layers and restores benign features at later layers, improving jailbreak resistance while preserving utility. Overall, these results suggest that pruning can support safety only when the intervention is explicitly trust-aware. Otherwise, compression may remove precisely the components that sustain aligned behavior.

Fairness Fairness-oriented pruning is less clearly aligned with inference-time trustworthiness than safety-oriented runtime pruning. Existing work mainly studies post-processing structural interventions, such as pruning attention heads associated with bias or comparing head- and neuron-level pruning for bias mitigation [173, 108]. These results suggest that selective structural suppression can reduce bias, but they are better viewed as adjacent pruning-based debiasing methods than as clean inference-time interventions. As a result, fairness remains relatively underexplored for pruning under the stricter inference-time setting considered here.

Privacy Pruning has also been explored for privacy protection, though much of this work is closer to deployment-oriented structural intervention than to strict runtime control during generation. Existing methods identify and remove neurons or subnetworks associated with memorized sensitive content, aiming to suppress privacy leakage while preserving general utility. Examples include data-efficient neuron pruning in language models, modality-aware pruning in MLLMs, and federated pruning schemes that exchange pruning masks rather than raw data [123, 106, 48]. These results suggest that pruning can support privacy-preserving deployment, but most current methods rely on structural modification before deployment rather than conditional pruning at inference time.

Factuality and Faithfulness Pruning has also shown some potential for improving factual reliability, although evidence remains limited. One large summarization study reports that pruned models hallucinate less and rely more on source content, suggesting that pruning may suppress spurious generation pathways [23]. However, it is still unclear whether this effect transfers beyond summarization or supports stronger notions of faithfulness. At present, factuality and faithfulness remain promising but underexplored directions for pruning in trustworthy model deployment.

4 System-Level Orchestration

System-level orchestration moves beyond single-model interventions to coordinate multiple LLM agents into collaborative architectures. Rather than controlling one model in isolation, trustworthiness here emerges from structured interaction patterns—debate, cross-verification, role specialization, and iterative self-correction. In the inference-time pipeline (Figure 1), the tool-use / agents loop spans context assembly, generation, and output checking, creating feedback cycles that enable collective reliability.

4.1 Multi-Agent Systems

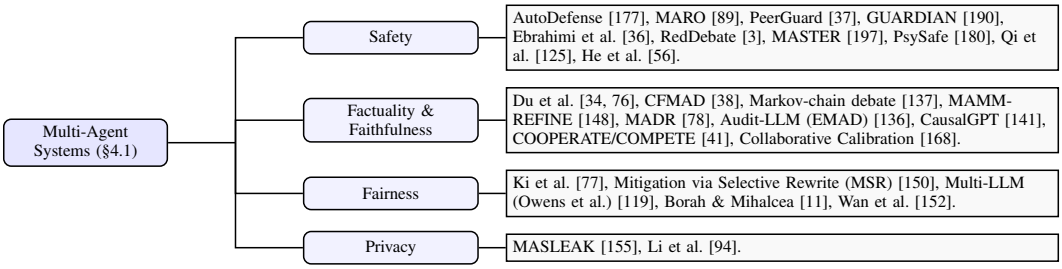

Figure 11: A taxonomy of Multi-Agent Systems for enhancing LLM trustworthiness.

Multi-agent systems (MAS) shift control from a single model to a coordinated group of LLM-based agents. Trustworthiness is no longer a property enforced on one model; it emerges from the interaction protocols that connect the agents. By structuring those interactions through debate, adversarial simulation, or multi-perspective deliberation, MAS can reach levels of robustness, safety, and factuality that single agents struggle to attain.

4.1.1 Core Principle: From Monolithic Control to Collaborative Systems

Multi-agent systems decentralize reasoning and decision-making. A complex problem is decomposed and assigned to specialized agents, each with its own role, persona, or expertise. Trustworthiness is shaped by the interaction protocol, which may use adversarial debate, cooperative problem-solving, peer review, or hierarchical verification. For example, a "propose" agent generates an initial response, a "critic" agent challenges its assumptions, and a "synthesize" agent integrates the feedback into a final answer. This structured interaction adds error correction and validation at inference time, so reliability depends less on any single agent's quality and more on the collaborative architecture itself.

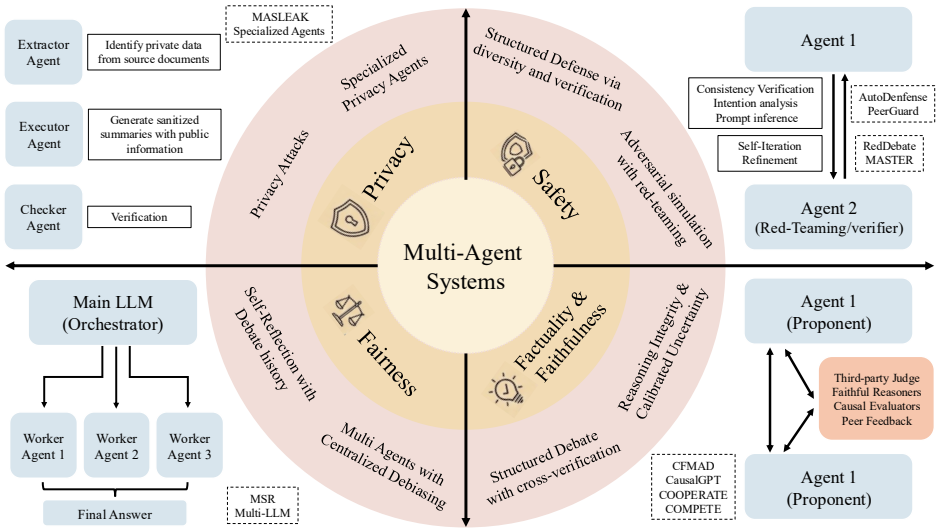

Figure 12: Overview of multi-agent system architectures for LLM trustworthiness across four dimensions. Safety employs structured defense via role diversity and adversarial red-teaming; Factuality relies on structured debate with cross-verification and calibrated uncertainty; Fairness uses multi-agent debiasing with self-reflection as well as orchestrated plan-and-refine loops; and Privacy leverages specialized agents (Extractor, Executor, Checker) for data sanitization. Representative systems and interaction patterns are shown for each dimension.

4.1.2 Applications

Safety Multi-agent systems make safety an emergent property of *structured interaction* rather than a static property of a single model. Specialized agents run a propose–attack–audit–adjudicate loop: they generate responses, probe for failures, cross-check reasoning, and arbitrate outcomes. A unifying pattern is *structured defense via role diversity and cross-verification*. AutoDefense combines intention analysis, prompt inference, and final judgment to filter unsafe outputs [177]. Committee-style analyzers extend coverage by combining linguistic, comment, and fact-checking expertise for misinformation risks [89]. Peer-based scrutiny hardens internal logic: PeerGuard has agents verify consistency between each other's chain-of-thought and final outputs to expose backdoors [37]. At system scale, collaboration can be modeled as a temporal interaction graph to trace and contain the spread of hallucinations or injected errors [190], while dynamic credibility weighting down-ranks unreliable contributors [36]. Safety also benefits from *adversarial simulation for proactive discovery*: automated red-teaming with debate iteratively elicits, refines, and patches unsafe behaviors using persistent memory [3]. The same interactive channels, however, expand the attack surface. Structured debates can amplify jailbreak success [125], communication links invite Agent-in-the-Middle manipulation [56], and persuasive agents can steer group consensus toward unsafe actions [180]; taxonomies such as MASTER map these vulnerabilities [197], and analyses of *emergent social risks* show how misaligned agent interactions can produce system-level harms that no individual agent was designed to cause [64].

Factuality Multi-agent systems enhance factuality and faithfulness by replacing a single model's unverified reasoning with structured interaction among multiple agents. Rather than trusting one model's first-pass answer, MAS organize a propose–critique–adjudicate loop in which agents advance claims, attack each other's reasoning, and converge under a judge. Foundational work shows that debating agents can reach more accurate conclusions than individuals, as critiques expose faulty logic that a single model may miss [34]; debates judged by weaker raters still tend to favor truthful arguments [76]. Forced-disagreement variants such as CFMAD assign counterfactual stances to improve robustness to initial mistakes [38]. This pattern scales to long-form tasks by decomposing outputs into atomic claims and verifying each through coordinated agents: role-structured stateful debates [137], discriminative reranking pipelines [148], complementary-perspective feedback for explanation faithfulness [78], and domain-specific evidence cross-examination [136]. Beyond surface correctness, MAS also promote reasoning integrity by separating faithful reasoners from causal evaluators [141], and yield calibrated uncertainty through peer feedback and adversarially generated alternatives that reveal knowledge gaps and trigger abstention [41, 168].

Fairness Multi-agent systems address demographic biases that simpler mitigation methods leave intact. One direction uses multi-agent debate for cultural adaptability: Ki et al. [77] propose a framework where LLM agents handle culturally situated questions through debate and self-reflection, with a judge LLM resolving disagreements based on debate history. For demographic fairness, multi-agent architectures decompose the debiasing task across specialized roles: Mitigation via Selective Rewrite uses an LLM agent to give targeted feedback on removing gender bias in language style [150]; multi-LLM frameworks place multiple models in conversational settings with centralized and decentralized debiasing configurations [119]; and self-reflection mechanisms let models identify and self-correct societal biases through a critique step [11]. Most recently, Wan et al. [152] proposes an agentic planning pipeline that runs an iterated diagnose-and-refine loop until the output is satisfactory.

Privacy Multi-agent systems introduce mechanisms for safety and factuality, but they also create new privacy challenges that remain a research gap. The interconnected structure of MAS opens new vectors for data leakage: the MASLEAK framework shows systematic, worm-like attacks that extract intellectual property such as system prompts, tool definitions, and topological structure from black-box MAS applications by propagating through agent communication chains [155]. The same architectural principles can also be used for defense: multi-agent approaches decompose privacy preservation into specialized roles, such as an Extractor Agent to identify private data, an Executor Agent to generate sanitized outputs, and a Checker Agent for verification, reducing leakage of sensitive information while preserving output quality [94]. Together, these studies show that the collaborative structure of MAS is both a liability and an asset for privacy. A unified treatment that

combines robust defenses, secure communication protocols, and verifiable information-flow controls remains largely unexplored.

5 Evaluation

Inference-time trustworthiness methods intervene after model training to shape LLM behavior under deployment constraints. As these interventions differ a lot in where and how they act, such as in prompts, logits, activations, external filters, or the system workflow, evaluation cannot be unified by one benchmark or one metric family. Although foundational benchmarks such as TrustGPT [62] and dynamic evaluation toolkits such as TrustEval [160] provide valuable unified views over multiple trustworthiness dimensions, they were not designed to isolate the runtime-specific properties introduced by inference-time controls. We therefore adopt a meta-axis evaluation view on the inference-time evaluation. The goal is to capture deployment-relevant properties that arise specifically from runtime trustworthiness. In practice, such enforcement is shaped by latency, cost, and integration constraints. Each axis highlights a distinct evaluation question, and while the axes are conceptually unique, they can overlap in practice:

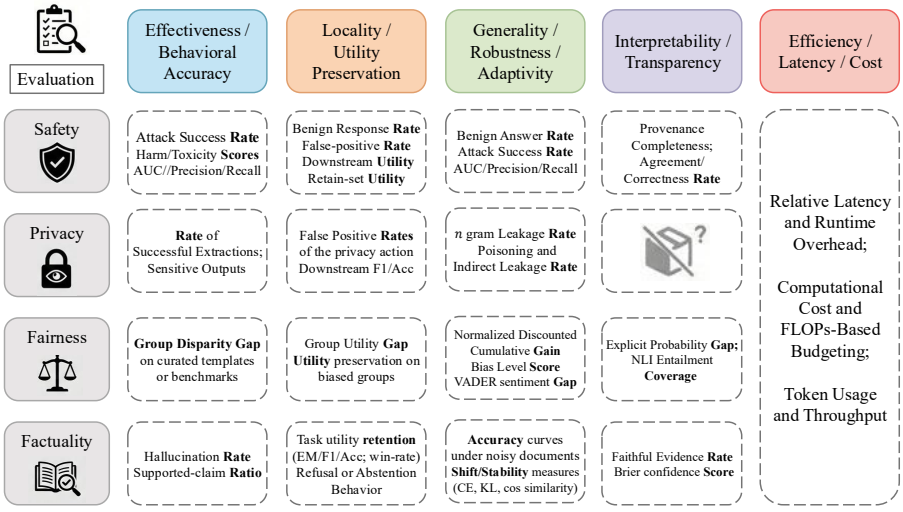

Figure 13: A meta-axis evaluation framework for inference-time trustworthiness methods. Rows correspond to four trustworthiness dimensions (Safety, Privacy, Fairness, Factuality), and columns represent five complementary evaluation axes (Effectiveness, Locality, Generality, Interpretability, Efficiency). Each cell lists representative metrics specific to the dimension–axis intersection, highlighting the multi-faceted nature of deployment-time evaluation.

- **Effectiveness / Behavioral Accuracy**: To what extent does the inference-time mechanism reliably enforce the intended behavioral constraints, and under what conditions does it fail?
- **Locality / Utility Preservation**: To what extent are the intervention effects localized to the targeted behaviors while preserving general task performance and user utility?
- **Generality / Robustness / Adaptivity**: How well does the mechanism generalize across tasks, domains, and input distributions, and how robust is it to distributional shifts?
- **Efficiency / Latency / Cost**: What additional runtime overhead does the mechanism introduce in terms of latency, compute, memory, or token usage?
- **Interpretability / Transparency**: To what extent are the mechanism's actions, triggers, and decision rationales transparent to human auditors or downstream systems?

Although these axes separate different deployment concerns, they often trade off against one another. Building on this framing, this section organizes evaluation along the axes above and describes representative metrics across four categories: Safety, Privacy, Fairness, and Factuality.

5.1 Effectiveness / Behavioral Accuracy

Effectiveness / Behavioral Accuracy evaluates whether an inference-time mechanism $\mathcal{M}$ changes model behavior in the intended direction on a target evaluation distribution. Consider an LLM that produces an output $Y_{\mathcal{M}}(x)$ for an input $x \sim \mathcal{D}$ under $\mathcal{M}$. Effectiveness asks whether $\mathcal{M}$ increases the likelihood of desired trust-aligned behavior and decreases the likelihood of undesired behavior:

$$\text{Effectiveness}(\mathcal{M}) \propto \mathbb{E}_{x\sim\mathcal{D}}\Big[\mathbf{1}\{Y_{\mathcal{M}}(x) \in \mathcal{Y}_{\text{desired}}\} - \mathbf{1}\{Y_{\mathcal{M}}(x) \in \mathcal{Y}_{\text{undesired}}\}\Big].$$

In inference-time settings, this axis primarily measures the runtime behavior induced by a plug-in control layer, such as refusal, redaction, constrained decoding, evidence grounding, or verification. However, Effectiveness alone does not tell whether improvements come from overly broad suppression of benign behavior, from brittle operating thresholds, or from masking the underlying capability. These trade-offs motivate complementary axes such as Locality (utility and over-refusal), Robustness (bypassability and shift), Efficiency (latency/cost), and Interpretability (auditability).

5.1.1 Rate-Based Metrics

Rate-based metrics capture how often undesired behavior still occurs at inference time, and are the most prevalent family across all trustworthiness categories. In safety, jailbreak success rate measures whether an adversarial prompt elicits a harmful response from the model, unsafe completion rate tracks the overall proportion of harmful outputs across all queries, and the complementary defense success rate captures how often the mechanism successfully blocks such outputs. Risk-score formulations take a continuous toxicity or harm score assigned to each output instead of a binary attack outcome and report the fraction exceeding a predefined threshold [80, 166, 59, 17, 27, 55, 37, 167]. In privacy, extraction success rate captures whether an attacker can recover memorized or sensitive text from model outputs, while disclosure rate measures how often outputs contain personally identifying or otherwise private information across a broader query distribution [122, 12, 154, 200]. In factuality, hallucination rate counts the proportion of outputs that are unsupported or factually incorrect, groundedness rate measures the fraction of claims in a response that are explicitly supported by retrieved or provided evidence, and edit success rate verifies whether a targeted fact has been correctly updated following a knowledge editing intervention [30, 4, 193, 111, 112].

5.1.2 Classifier-Style Metrics

Classifier-style metrics treat the inference-time intervention as an explicit detection task and evaluate how well the mechanism separates harmful, private, or hallucinated outputs from benign ones. Rather than reporting a single operating point, metrics such as AUC and AUPRC summarize performance across all possible decision thresholds, while Precision, Recall, and F1 show the trade-off between false positives and false negatives at a chosen threshold. These metrics are important when the intervention involves a learned or rule-based classifier, such as a safety guardrail or hallucination detector, where threshold calibration governs deployment behavior. They appear across safety guardrails [120, 22, 82, 103], privacy membership-inference evaluations [200, 116], and factuality hallucination detection [4, 193].

5.1.3 Group-Disparity Metrics

Group-disparity metrics quantify fairness effectiveness by comparing outcomes across demographic groups, capturing whether a mechanism treats different populations equitably. Bias scores on curated benchmarks measure stereotyping or differential toxicity for specific groups, worst-group performance identifies the most disadvantaged subpopulation, and disparity gaps capture the absolute difference in an outcome metric such as toxicity, error rate, or refusal rate between two groups. Together, these metrics reveal whether a fairness intervention reduces imbalance in model behavior or merely shifts it. They are reported across prompt-based interventions [35, 60, 74], fairness-aware retrieval [21, 135], and pruning methods that target bias-related internal structure [14, 173, 108].

5.1.4 Task-Performance Metrics

Task-performance metrics such as exact match, F1, and accuracy measure whether trustworthiness gains come at the cost of general capability, serving as a sanity check on standard benchmarks. They

are widely reported among representation or context engineering evaluations to confirm that they do not disrupt unrelated competencies [91, 7, 157, 111], and also appear in multimodal hallucination settings [121, 172].

5.2 Locality / Utility Preservation

Locality / Utility Preservation evaluates whether an inference-time mechanism $\mathcal{M}$ limits its behavioral changes to the intended outputs while preserving user utility. Unlike training-time alignment, inference-time controls are typically implemented as plug-in layers. As a result, the locality failure mode is behavioral spillover, in which the system becomes safer by being less helpful, issuing more refusals, or providing less information. Formally, let $\mathcal{D}_{\text{tgt}}$ denote the targeted distribution (e.g., harmful or sensitive prompts) and $\mathcal{D}_{\text{benign}}$ denote benign traffic. A generic utility retention view is:

$$\Delta U_{\text{benign}}(\mathcal{M}) = \mathbb{E}_{x \sim \mathcal{D}_{\text{benign}}}[u(Y_{\mathcal{M}}(x)) - u(Y(x))]\,,$$

where $u(\cdot)$ can be task accuracy (EM/F1/Acc), LLM-as-judge helpfulness, or win-rate. In inference-time deployment, locality is coupled to thresholding and pipeline composition. When multiple strategies are combined, such as guardrails and decoding, they can amplify suppression of benign outputs even if each component appears acceptable in isolation.

5.2.1 False-Positive and Over-Refusal Rate Metrics

These metrics capture whether a mechanism incorrectly suppresses or refuses inputs it was not designed to target. In safety, benign response rate measures how often the defense leaves safe prompts unaffected, and is usually reported alongside attack success rate to confirm that refusal gains are not simply the result of refusing more broadly [80, 15, 68, 194]. In guardrail systems, false-positive rate and benign blocking rate capture how often clean inputs are incorrectly filtered or routed. These summarize the trade-off between blocking harmful content and passing benign traffic [120, 22, 82, 103]. In multi-agent defenses, false alarm rates on clean interactions are essential because a monitoring agent that incorrectly flags benign exchanges can suppress the entire downstream workflow [37, 3]. In privacy, false positives of the privacy action through blocking harmless queries, measure damage on non-sensitive content [122, 154]. In fairness, refusal disparity measures the absolute difference in refusal rate between demographic groups under the intervention, capturing whether the mechanism disproportionately suppresses responses for certain groups rather than improving equitable helpfulness [35, 60, 74].

5.2.2 Utility Retention Metrics

Utility retention metrics measure whether general task performance is preserved on non-targeted inputs after the intervention. Common metrics include exact match, F1, and accuracy on held-out benchmarks, MT-Bench scores, instruction-following rates, LLM-as-judge helpfulness, and win-rate. In safety, these are reported for decoding, representation engineering, pruning, and unlearning interventions to confirm that modifications do not distort normal generation quality [166, 2, 17, 55, 27, 181]. In privacy, downstream task accuracy on non-sensitive queries confirms that filtering or constrained generation does not broadly reduce answer quality [12, 200]. In fairness, per-group utility retention tracks whether task success is preserved equitably across demographic groups, ensuring bias reduction is not obtained by broadly degrading helpfulness for any subpopulation [135, 14, 173, 108]. In factuality, retain-set task utility on benign factual QA is reported alongside hallucination reductions [30, 91, 4, 193].

5.3 Generality / Robustness / Adaptivity

Generality / Robustness / Adaptivity evaluate whether an inference time mechanism $\mathcal{M}$ keeps its intended effect when inputs, tasks, or attackers change at deployment. This axis is not about a single test set. It is about performance over a family of conditions, which is likely out-of-distribution. Let $\{\mathcal{D}_k\}_{k=1}^{K}$ be test distributions and let $s_k(\mathcal{M})$ be a score on $\mathcal{D}_k$. A common summary reports both the average and the worst case:

$$s_{\text{avg}}(\mathcal{M}) = \frac{1}{K}\sum_{k=1}^{K} s_k(\mathcal{M}), \qquad s_{\min}(\mathcal{M}) = \min_{k \in [K]} s_k(\mathcal{M}).$$

5.3.1 Worst-Case and Cross-Distribution Rate Metrics

These metrics summarize how a mechanism's core effectiveness signal holds up across multiple conditions rather than a single fixed distribution. In safety, attack success rate and benign answering rate are reported across multiple jailbreak families and prompt styles, with worst-case values serving as the primary robustness summary. Sensitivity to threshold changes is captured by reporting deltas when guard strength varies [80, 15, 194]. Decoding and representation engineering methods report harmful score and safety score under stronger or more diverse attacks. Tracking utility accuracy on task suites to confirm that robustness gains do not come at the cost of normal performance [166, 156, 27]. In privacy, attack success rate is reported across multiple attack types and privacy budgets to capture how leakage behavior changes as the adversary changes. N-gram leakage rate measures whether the control generalizes beyond exact span matches to broader overlap with sensitive content [143, 16]. In fairness, bias scores and stereotype index are reported across demographic categories, topics, and occupations to evaluate whether bias reduction holds beyond a fixed template. Fairness violation rates are tracked alongside utility metrics such as NDCG [74, 35, 60]. In factuality, accuracy and F1 are reported across cross-domain splits to test whether robustness comes from better evidence selection rather than longer contexts [149, 30, 193, 4].

5.3.2 Classifier-Style Metrics Across Domains and Languages

When the intervention involves an explicit detector or classifier, robustness is evaluated by reporting precision, recall, F1, AUC, and AUPRC across multiple benchmarks, domains, and languages. In safety, guardrails and multi-agent systems use these metrics to evaluate whether detection quality generalizes across domains and languages [120, 46, 3]. In privacy, precision, recall, and F1 for detecting poisoned passages evaluate whether the control generalizes to different poisoning strategies, since both false positives and false negatives matter [200, 188]. In fairness, detailed metrics such as sentiment gap, toxicity scores, and regard-based labels are reported across demographic groups to evaluate how robust the fairness conclusion is to different measurement approaches [173, 25].

5.3.3 Calibration and Stability Metrics

These metrics evaluate whether a mechanism remains consistent and reliable as conditions shift. In safety, expected calibration error tests whether the same decision threshold remains reliable under distribution shift. In factuality, representation engineering methods track shift measures such as cross-entropy and KL divergence to detect distribution change, and stability measures such as cosine similarity to check whether the intervention direction stays consistent as training data changes [91, 198, 7]. In fairness, cross-split and cross-capability comparisons, reporting generality scores on train-test splits, check whether bias reductions are stable rather than artifacts of a particular evaluation setup [126, 14].

5.4 Interpretability / Transparency

Interpretability / Transparency evaluates whether an inference-time mechanism $\mathcal{M}$ is observable and auditable. When $\mathcal{M}$ changes the system behavior (e.g., refusal, redaction, or rewriting), it should be clear (i) what action was taken, (ii) what condition triggered the action, and (iii) what evidence supports the decision. This is particularly important at inference time because many controls are deployed as modular plug-ins. Formally, let $a_{\mathcal{M}}(x) \in \mathcal{A}$ denote the runtime action on input x, and let $\tau_{\mathcal{M}}(x)$ be the trigger set (policy IDs, detectors, or threshold crossings). Let $z_{\mathcal{M}}(x)$ be the transparency artifact (logs, scores, or traces). A generic transparency score is:

$$T(\mathcal{M}) = \mathbb{E}_{x\sim\mathcal{D}}\left[v(a_{\mathcal{M}}(x),\, \tau_{\mathcal{M}}(x),\, z_{\mathcal{M}}(x))\right],$$

where $s(x)$ is a risk score (or classifier confidence) and t is the deployed threshold.

5.4.1 Provenance and Audit Metrics

These metrics measure whether each intervention is accompanied by a sufficient audit record that identifies what component acted, what policy triggered it, and at what operating point. In safety, provenance completeness captures whether each intervention includes a stage identifier, policy identifier, risk score, and threshold value, making it possible to distinguish targeted filtering from broad over-refusal [136, 190, 103, 17]. In factuality, provenance logs of what was retrieved and

used at runtime help separate retrieval failures from generation failures, and verification loops or critique-style retrieval augmentation leave intermediate records that make the reasoning process auditable [30, 4, 149].

5.4.2 Human Judgment Metrics

These metrics evaluate whether the mechanism's explanations are understandable to human annotators. In safety, agreement rate measures the fraction of explanations that annotators judge as correct or helpful, evaluating whether intermediate reasoning steps and policy triggers are meaningful rather than superficially plausible [68, 75]. In factuality, faithful explanation rate measures the fraction of explanations that humans consider consistent with the provided evidence, evaluating whether verification rationales are genuinely grounded [78, 4]. LLM-as-a-judge is widely used as a scalable proxy for human judgment, but recent work shows that judge models themselves carry systematic biases—e.g., position, verbosity, and self-preference biases—which can distort transparency evaluations and should be reported alongside raw judge scores [169].

5.4.3 Distributional Transparency Metrics

These metrics evaluate whether the mechanism's behavior across groups or confidence levels is auditable from the output distribution. In fairness, an explicit probability gap between two matched identity terms in the same context measures how much the mechanism shifts likelihoods across groups [14]. In pluralistic settings, NLI entailment coverage measures how much the final response reflects provided group viewpoints, capturing whether the mechanism genuinely incorporates intended perspectives [126]. In factuality, calibration metrics such as the Brier score measure whether reported confidence aligns with actual correctness, making the system's factuality claims auditable across varying operating points [76, 168].

5.5 Efficiency / Latency / Cost

Efficiency / Latency / Cost evaluates how practical it is to deploy an inference-time mechanism $\mathcal{M}$ by measuring the computational resources consumed during the control process. Unlike training-time alignment, which spreads costs over the model's lifetime, inference-time methods operate on the critical path of user interaction. Even small increases in latency or compute per request can noticeably worsen user experience and raise operational costs. The main efficiency failure mode is excessive overhead, where the mechanism requires many extra forward passes, memory, or auxiliary calls to achieve safety or accuracy improvements. Formally, let $C_{\text{base}}(x)$ denote the cost of generating a response for input x using the base model. A general overhead measure is:

$$\Delta C(\mathcal{M}) = \mathbb{E}_{x\sim\mathcal{D}}[C_{\mathcal{M}}(x) - C_{\text{base}}(x)]\,,$$

where $C_{\mathcal{M}}(x)$ is the cost under the mechanism. The commonly used metrics are:

- **Relative Latency and Runtime Overhead.** A standard way to measure deployment cost is relative latency, often reported as the *Average Token Generation Time Ratio* or its equivalence [166, 101, 96]. It measures the slowdown caused by the defense during inference:

  $$\text{ATGR} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{\text{def}}(x_i)}{T_{\text{base}}(x_i)},$$

  where $T_{\text{def}}(x_i)$ is the wall-clock time for prompt x_i under the defense, and $T_{\text{base}}(x_i)$ is the baseline time.

- **Computational Cost and FLOPs-Based Budgeting.** To study scaling behavior in a way that is less sensitive to hardware, several works use floating point operations as a compute proxy [127, 86]. This gives a hardware-independent view of how much extra computation is needed for a given safety gain, often reported as performance under a compute budget:

  $$\text{Performance}(C_{\text{budget}}) = \text{Metric}(\mathcal{M} \mid C_{\mathcal{M}} \le C_{\text{budget}})\,.$$

  This is used to plot safety outcomes such as Attack Success Rate against inference TFLOPs, tracing a safety–compute trade-off curve.

- **Token Usage and Throughput.** For methods that expand context—such as guardrail prompting, meta-prompting, or retrieval augmentation—token growth is a direct cost driver [69, 135, 176]. A simple and widely used summary is the *Token Usage Ratio*:

$$r_{TT} = \frac{\text{Total Tokens}_{\text{defense}}}{\text{Total Tokens}_{\text{base}}}.$$

This captures cost increases from longer prompts, extra intermediate steps, or additional retrieved content. Throughput in tokens per second is also commonly reported to reflect serving capacity under the defense.

6 Discussion and Open Problems

6.1 External Controls

Context engineering, guardrails, and decoding strategies share a common strength: they treat the model as a black box, making them model-agnostic, composable, and rapidly updatable without retraining. Context engineering is training-free and flexible, with different techniques addressing distinct challenges: prompting for lightweight alignment, RAG for factual grounding, memory for persistence, and reasoning for self-regulation. Guardrails add an independent policy-enforcement layer that can be swapped across providers, while decoding strategies enable fine-grained, real-time steering of token distributions across multiple trustworthiness dimensions.

However, all three tiers are soft controls and share similar failure modes. Context engineering is fragile to small phrasing changes and vulnerable to context poisoning; guardrails face a continuous cat-and-mouse dynamic against adversarial evasion; and hard decoding constraints can produce brittle, incoherent outputs [59]. A common "Control Tax" runs through these methods: strict guardrails increase false positives and reduce helpfulness, while aggressive logit suppression [183] or noise injection [109] compromises generation quality. LLM-based judges and multi-layer decoding also add latency overhead. None of these methods can provide formal safety guarantees in isolation, and integrated pipelines that combine context engineering, guardrails, and decoding remain underexplored.

6.2 Internal Manipulations

Representation engineering, unlearning, and pruning offer more direct behavioral control by intervening on internal model components. Representation engineering treats hidden states as controllable objects and supports concept-level steering through activation addition or sparse autoencoders [198, 91, 144]. Unlearning steers outputs away from sensitive knowledge through runtime penalties or influence subtraction, which is practical when regulatory requirements shift. Pruning provides structural control by removing heads, neurons, or tokens associated with unsafe or biased behavior, and is compatible with broader compression pipelines.

Despite these strengths, internal methods share several fragilities. Representation engineering depends heavily on intervention strength, layer choice, and normalization; multi-concept steering remains brittle, and strong interventions risk pushing activations out of distribution [161]. Unlearning faces a verification problem: many methods suppress rather than truly remove knowledge, leaving models vulnerable to adversarial recovery [107]. Pruning is highly sensitive to where sparsity is introduced; even small amounts of naive pruning can sharply weaken safety-critical behavior. All three methods exhibit a utility–control trade-off, where stronger interventions can degrade fluency, accuracy, or reasoning [104]. Representation engineering also carries dual-use risks, as the same activation-level controls can strengthen or weaken safety [93]. Robust deployment across internal methods will require better verification protocols, principled intervention selection, and careful integration with broader safety governance [146].

6.3 System-Level Orchestration

Multi-agent systems introduce reliability through collaborative reasoning, but their most immediate obstacle is operational overhead: each query may trigger multiple agents over several rounds, with performance gains often diminishing quickly [38, 34], which leaves many architectures impractical for real-time applications [136]. The collaborative dynamics designed to foster trustworthiness

can also become vectors for emergent failures. Agents sharing the same base model exhibit "groupthink," reinforcing shared hallucinations [38], and adversaries can exploit the collaborative structure: a single persuasive adversary can sway an entire group [1], while debate frameworks can be more vulnerable to jailbreaks than single agents [125]. Evaluating MAS trustworthiness is also more complex, as effectiveness depends on roles, topology, protocols, and agent count [197].

7 Conclusion

As large language models are deployed in real applications, controlling their trustworthiness after training has become as important as alignment during training. In this paper, we have presented a unified framework for inference-time control of trustworthy LLM behavior across safety, privacy, fairness, and factuality. The framework spans external guardrails, internal representation engineering, and system-level orchestration. Our analysis shows that no single method is sufficient; each trades off control strength, computational cost, and model utility against the others. Building reliable systems will therefore require composing these methods rather than choosing among them. Inference-time intervention is not an afterthought to alignment but a coherent design space in its own right, and treating it as such is a starting point for verifiably trustworthy LLMs.

References

[1] Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6929–6948, 2024.

[2] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

[3] Ali Asad, Stephen Obadinma, Radin Shayanfar, and Xiaodan Zhu. Reddebate: Safer responses through multi-agent red teaming debates, 2025.

[4] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Somnath Banerjee, Soham Tripathy, Sayan Layek, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *AAAI Conference on Artificial Intelligence*, 2024.

[6] Fazl Barez, Tingchen Fu, Ameya Prabhu, Stephen Casper, Amartya Sanyal, Adel Bibi, Aidan O'Gara, Robert Kirk, Ben Bucknall, Tim Fist, et al. Open problems in machine unlearning for ai safety. *URL https://arxiv. org/abs/2501.04952*, 2025.

[7] Farima Fatahi Bayat, Xin Liu, H. V. Jagadish, and Lu Wang. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

[8] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*, 2023.

[9] Baolong Bi, Shenghua Liu, Lingrui Mei, Yiwei Wang, Pengliang Ji, and Xueqi Cheng. Decoding by contrasting knowledge: Enhancing llms' confidence on edited facts. *ArXiv*, abs/2405.11613, 2024.

[10] Rob Bonta. California consumer privacy act (ccpa). *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa*, 2022.

[11] Angana Borah and Rada Mihalcea. Towards implicit bias detection and mitigation in multi-agent LLM interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9306–9326, 2024.

[12] Luka Borec, Philipp Sadler, and David Schlangen. The unreasonable ineffectiveness of nucleus sampling on mitigating text memorization. In *International Conference on Natural Language Generation*, 2024.

[13] Daniel Borkan, Lucas Dixon, Jeffrey Scott Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[14] Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. Locating and mitigating gender bias in large language models. In *International Conference on Intelligent Computing*, pages 471–482, 2024.

[15] Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. Defending against alignment-breaking attacks via robustly aligned llm. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 10542–10560, 2024.

[16] Harsh Chaudhari, Giorgio Severi, John Abascal, Matthew Jagielski, Christopher A Choquette-Choo, Milad Nasr, Cristina Nita-Rotaru, and Alina Oprea. Phantom: General trigger attacks on retrieval augmented language generation. *arXiv preprint arXiv:2405.20485*, 2024.

[17] Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. *arXiv preprint arXiv:2507.01513*, 2025.

[18] Bocheng Chen, Hanqing Guo, and Qiben Yan. Flexllm: Exploring llm customization for moving target defense on black-box llms against jailbreak attacks. *ArXiv*, abs/2412.07672, 2024.

[19] Dingwei Chen, Feiteng Fang, Shiwen Ni, Feng Liang, Xiping Hu, Ahmadreza Argha, Hamid Alinejad-Rokny, Min Yang, and Chengming Li. Lower layers matter: Alleviating hallucination via multi-layer fusion contrastive decoding with truthfulness refocused. 2024.

[20] Mingda Chen, Yang Li, Karthik Padthe, Rulin Shao, Alicia Sun, Luke Zettlemoyer, Gargi Ghosh, and Wen-tau Yih. Improving factuality with explicit working memory. *arXiv preprint arXiv:2412.18069*, 2024.

[21] Quan Ze Chen, KJ Feng, Chan Young Park, and Amy X Zhang. Spica: Retrieving scenarios for pluralistic in-context alignment. *arXiv preprint arXiv:2411.10912*, 2024.

[22] Jianfeng Chi, Ujjwal Karn, Hongyuan Zhan, Eric Michael Smith, Javier Rando, Yiming Zhang, Kate Plawiak, Zacharie Delpierre Coudert, K. Upasani, and Mahesh Pasupuleti. Llama guard 3 vision: Safeguarding human-ai image understanding conversations. *ArXiv*, abs/2411.10414, 2024.

[23] George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181, 2024.

[24] M. Das and Wolf-Tilo Balke. Quantifying bias from decoding techniques in natural language generation. In *International Conference on Computational Linguistics*, 2022.

[25] Vishnu Asutosh Dasu, Vipul Gupta, Saeid Tizpaz-Niari, Gang Tan, et al. Attention pruning: Automated fairness repair of language models via surrogate simulated annealing. *arXiv preprint arXiv:2503.15815*, 2025.

[26] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*, 2017.

[27] Zhijie Deng, Chris Yuhao Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Jiaheng Wei. Guard: Generation-time llm unlearning via adaptive restriction and detection. *arXiv preprint arXiv:2505.13312*, 2025.

[28] J. Dhamala, Varun Kumar, Rahul Gupta, Kai-Wei Chang, and A. G. Galstyan. An analysis of the effects of decoding algorithms on fairness in open-ended language generation. *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 655–662, 2022.

[29] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*, 2023.

[30] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. *FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: ACL 2024*, pages 3563–3578, 2024.

[31] Leo Donisch, Sigurd Schacht, and Carsten Lanquillon. Inference optimizations for large language models: Effects, challenges, and practical considerations. *arXiv preprint arXiv:2408.03130*, 2024.

[32] Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. Avoiding copyright infringement via large language model unlearning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5176–5200, 2025.

[33] Yanrui Du, Sendong Zhao, Danyang Zhao, Ming Ma, Yuhan Chen, Liangyu Huo, Qing Yang, Dongliang Xu, and Bing Qin. Mogu: A framework for enhancing safety of open-sourced llms while preserving their usability. *ArXiv*, abs/2405.14488, 2024.

[34] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[35] Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. Breaking the bias: Gender fairness in llms using prompt engineering and in-context learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4), 2023.

[36] Sana Ebrahimi, Mohsen Dehghankar, and Abolfazl Asudeh. An adversary-resistant multi-agent llm system via credibility scoring, 2025.

[37] Falong Fan and Xi Li. Peerguard: Defending multi-agent systems against backdoor attacks through mutual reasoning, 2025.

[38] Yi Fang, Moxin Li, Wenjie Wang, Lin Hui, and Fuli Feng. Counterfactual debating with preset stances for hallucination elimination of llms. In *COLING*, pages 10554–10568, 2025.

[39] Arya Fayyazi, Mehdi Kamal, and Massoud Pedram. FACTER: Fairness-aware conformal thresholding and prompt engineering for enabling fair LLM-based recommender systems. In *Forty-second International Conference on Machine Learning*, 2025.

[40] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. Teaching llms to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, 2024.

[41] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, 2024.

[42] Víctor Gallego. Metasc: Test-time safety specification optimization for language models. *ArXiv*, abs/2502.07985, 2025.

[43] Jiahui Gao, Renjie Pi, Tianyang Han, Han Wu, Lanqing Hong, Lingpeng Kong, Xin Jiang, and Zhenguo Li. Coca: Regaining safety-awareness of multimodal large language models with constitutional calibration. *ArXiv*, abs/2409.11365, 2024.

[44] Lang Gao, Jiahui Geng, Xiangliang Zhang, Preslav Nakov, and Xiuying Chen. Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25378–25398, 2025.

[45] Soumya Suvra Ghosal, Souradip Chakraborty, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, and A. S. Bedi. Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25038–25049, 2024.

[46] Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive ai content safety moderation with ensemble of llm experts. *ArXiv*, abs/2404.05993, 2024.

[47] Gene H. Golub and Christian Reinsch. Singular value decomposition and least squares solutions. *Linear Algebra*, pages 134–151, 1971.

[48] Pengxin Guo, Yinong Wang, Wei Li, Mengting Liu, Ming Li, Jinkai Zheng, and Liangqiong Qu. Exploring federated pruning for large language models. *arXiv preprint arXiv:2505.13547*, 2025.

[49] William Hackett, Lewis Birch, Stefan Trawicki, Neeraj Suri, and Peter Garraghan. Bypassing prompt injection and jailbreak detection in llm guardrails. *arXiv preprint arXiv:2504.11168*, 2025.

[50] Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. Safeswitch: Steering unsafe llm behavior via internal activation signals. *arXiv preprint arXiv:2502.01042*, 2025.

[51] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms, 2024.

[52] Shanshan Han, Amir Salman Avestimehr, and Chaoyang He. Bridging the safety gap: A guardrail pipeline for trustworthy llm inferences. *ArXiv*, abs/2502.08142, 2025.

[53] Sa hana Chennabasappa, Cyrus Nikolaidis, Daniel Song, David Molnar, Stephanie Ding, Shengye Wan, Spencer Whitman, Lauren Deason, Nicholas Doucette, Abraham Montilla, Alekhya Gampa, Beto de Paola, Dominik Gabi, James Crnkovich, Jean-Christophe Testud, Kat He, Rashnil Chaturvedi, Wu Zhou, and Joshua Saxe. Llamafirewall: An open source guardrail system for building secure ai agents. *ArXiv*, abs/2505.03574, 2025.

[54] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

[55] Adib Hasan, Ileana Rugina, and Alex Wang. Pruning for protection: Increasing jailbreak resistance in aligned LLMs without fine-tuning. In *The 7th BlackboxNLP Workshop*, 2024.

[56] Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. Red-teaming LLM multi-agent systems via communication attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6726–6747, 2025.

[57] Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

[58] Jakub Hościłowicz, Adam Wiacek, Jan Chojnacki, Adam Cieślak, Leszek Michon, Vitalii Urbanevych, and Artur Janicki. Non-linear inference time intervention: Improving llm truthfulness, 2024.

[59] Caishuang Huang, Wanxu Zhao, Rui Zheng, Huijie Lv, Shihan Dou, Sixian Li, Xiao Wang, Enyu Zhou, Junjie Ye, Yuming Yang, Tao Gui, Qi Zhang, and Xuanjing Huang. Safealigner: Safety alignment against jailbreak attacks via response disparity guidance. *ArXiv*, abs/2406.18118, 2024.

[60] Dong Huang, Jie M Zhang, Qingwen Bu, Xiaofei Xie, Junjie Chen, and Heming Cui. Bias testing and mitigation in llm-based code generation. *ACM Transactions on Software Engineering and Methodology*, 2024.

[61] Yanwen Huang, Yong Zhang, Ning Cheng, Zhitao Li, Shaojun Wang, and Jing Xiao. Dynamic attention-guided context decoding for mitigating context faithfulness hallucinations in large language models. In *Annual Meeting of the Association for Computational Linguistics*, 2025.

[62] Yue Huang, Qihui Zhang, Philip S. Yu, and Lichao Sun. TrustGPT: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023.

[63] Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*, 2024.

[64] Yue Huang, Yu Jiang, Wenjie Wang, Haomin Zhuang, Xiaonan Luo, Yuchen Ma, Zhangchen Xu, Zichen Chen, Nuno Moniz, Zinan Lin, Pin-Yu Chen, Nitesh V. Chawla, Nouha Dziri, Huan Sun, and Xiangliang Zhang. Emergent social intelligence risks in generative multi-agent systems. *arXiv preprint arXiv:2603.27771*, 2026.

[65] Yue Huang, Haomin Zhuang, Jiayi Ye, Han Bao, Yanbo Wang, Hang Hua, Siyuan Wu, Pin-Yu Chen, and Xiangliang Zhang. Guardian-as-an-advisor: Advancing next-generation guardian models for trustworthy llms. *arXiv preprint arXiv:2604.07655*, 2026.

[66] Hakan Inan, K. Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. Llama guard: Llm-based input-output safeguard for human-ai conversations. *ArXiv*, abs/2312.06674, 2023.

[67] Masaru Isonuma and Ivan Titov. What's new in my data? novelty exploration via contrastive generation. *ArXiv*, abs/2410.14765, 2024.

[68] Jiabao Ji, Bairu Hou, Alexander Robey, George J Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. Defending large language models against jailbreak attacks via semantic smoothing. *arXiv preprint arXiv:2402.16192*, 2024.

[69] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Boxin Wang, Jinyuan Jia, Bo Li, and Radha Poovendran. Poster: Identifying and mitigating vulnerabilities in llm-integrated applications. *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, 2023.

[70] Fengqing Jiang, Zhangchen Xu, Luyao Niu, Zhen Xiang, Bhaskar Ramasubramanian, Bo Li, and Radha Poovendran. Artprompt: Ascii art-based jailbreak attacks against aligned llms. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[71] Xinyan Jiang, Lin Zhang, Jiayi Zhang, Qingsong Yang, Guimin Hu, Di Wang, and Lijie Hu. Msrs: Adaptive multi-subspace representation steering for attribute alignment in large language models. *arXiv preprint arXiv:2508.10599*, 2025.

[72] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[73] Fanny Jourdan, Louis Béthune, Agustin Picard, Laurent Risser, and Nicholas Asher. Taco: Targeted concept erasure prevents non-linear classifiers from detecting protected attributes. *arXiv preprint arXiv:2312.06499*, 2024.

[74] Mahammed Kamruzzaman and Gene Louis Kim. Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. *arXiv preprint arXiv:2404.17218*, 2024.

[75] Mintong Kang and Bo Li. R2-guard: Robust reasoning enabled llm guardrail via knowledge-enhanced logical reasoning. *ArXiv*, abs/2407.05557, 2024.

[76] Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[77] Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. Multiple LLM agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877, 2025.

[78] Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate, 2024.

[79] Sahasra Kokkula, R Somanathan, R Nandavardhan, Aashishkumar, and G Divya. Palisade - prompt injection detection framework. *ArXiv*, abs/2410.21146, 2024.

[80] Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Aaron Jiaxun Li, Soheil Feizi, and Himabindu Lakkaraju. Certifying llm safety against adversarial prompting. In *First Conference on Language Modeling*, 2024.

[81] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[82] Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. Polyguard: A multilingual safety moderation tool for 17 languages. *ArXiv*, abs/2504.04377, 2025.

[83] Yicheng Lang, Kehan Guo, Yue Huang, Yujun Zhou, Haomin Zhuang, Tianyu Yang, Yao Su, and Xiangliang Zhang. Beyond single-value metrics: Evaluating and enhancing LLM unlearning with cognitive diagnosis. *arXiv preprint arXiv:2502.13996*, 2025.

[84] LangChain AI. Rebuff: Prompt injection detection for llm applications. `https://blog.langchain.com/rebuff/`, 2023. Accessed: 2025-08-18.

[85] Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2024.

[86] Seanie Lee, Dong Bok Lee, Dominik Wagner, Minki Kang, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. Saferoute: Adaptive model selection for efficient and accurate safety guardrails in large language models. *ArXiv*, abs/2502.12464, 2025.

[87] Alyssa Lees, Vinh Q. Tran, Yi Tay, Jeffrey Scott Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. A new generation of perspective api: Efficient multilingual character-level transformers. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[88] Haohang Li, Yupeng Cao, Yangyang Yu, Jordan W. Suchow, and Zining Zhu. Truth neurons, 2025.

[89] Hui Li, Ante Wang, kunquan li, Zhihao Wang, Liang Zhang, Delai Qiu, Qingsong Liu, and Jinsong Su. A multi-agent framework with automated decision rule optimization for cross-domain misinformation detection, 2025.

[90] Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, Nouha Dziri, Anne Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and Sydney Levine. Safetyanalyst: Interpretable, transparent, and steerable safety moderation for ai behavior. In *Forty-second International Conference on Machine Learning*, 2025.

[91] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024.

[92] Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11858–11875, 2024.

[93] Tianyu Li et al. Revisiting jailbreaking for large language models. In *COLING*, 2025.

[94] Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, and Maarten Sap. 1-2-3 check: Enhancing contextual privacy in LLM via multi-agent reasoning. In *Proceedings of the The First Workshop on LLM Security (LLMSEC)*, pages 115–128, 2025.

[95] Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. Fairsteer: Inference-time debiasing for llms with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312, 2025.

[96] Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. Cleangen: Mitigating backdoor attacks for generation tasks in large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

[97] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[98] Xiaoyu Liang, Jiayuan Yu, Lianrui Mu, Jiedong Zhuang, Jiaqi Hu, Yuchen Yang, Jiangnan Ye, Lu Lu, Jian Chen, and Haoji Hu. Mitigating hallucination in visual-language models via re-balancing contrastive decoding. In *Chinese Conference on Pattern Recognition and Computer Vision*, 2024.

[99] Jiacheng Liu, Peng Tang, Wenfeng Wang, Yuhang Ren, Xiaofeng Hou, Pheng-Ann Heng, Minyi Guo, and Chao Li. A survey on inference optimization techniques for mixture of experts models. *arXiv preprint arXiv:2412.14219*, 2024.

[100] Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. On the universal truthfulness hyperplane inside llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[101] Quan Liu, Zhenhong Zhou, Longzhu He, Yi Liu, Wei Zhang, and Sen Su. Alignment-enhanced decoding: Defending jailbreaks via token-level adaptive refining of probability distributions. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

[102] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*, 2023.

[103] Yue Liu, Hongcheng Gao, Shengfang Zhai, Jun Xia, Tianyi Wu, Zhiwei Xue, Yulin Chen, Kenji Kawaguchi, Jiaheng Zhang, and Bryan Hooi. Guardreasoner: Towards reasoning-based llm safeguards. *ArXiv*, abs/2501.18492, 2025.

[104] Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. Breaking the trilemma of privacy, utility, and efficiency via controllable machine unlearning. In *Proceedings of the ACM Web Conference 2024*, pages 1260–1271, 2024.

[105] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*, 2024.

[106] Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025.

[107] Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

[108] Sibo Ma, Alejandro Salinas, Julian Nyarko, and Peter Henderson. Breaking down bias: On the limits of generalizable pruning strategies. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2025.

[109] Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard S. Zemel. Differentially private decoding in large language models. *ArXiv*, abs/2205.13621, 2022.

[110] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *AAAI Conference on Artificial Intelligence*, 2020.

[111] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[112] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *ArXiv preprint*, abs/2210.07229, 2022.

[113] Meta. Llama prompt guard documentation. Meta (online), 2025. Model card and prompt-format documentation for Llama Prompt Guard, available online at Meta's official site.

[114] Microsoft. Prompt shields in azure ai content safety. Microsoft Learn (online), 2025. Accessed via Microsoft Learn documentation: unified API to detect and block adversarial user-input attacks on LLMs.

[115] Andrei Ioan Muresanu, Anvith Thudi, Michael R Zhang, and Nicolas Papernot. Fast exact unlearning for in-context learning data for llms. In *ICML*, 2025.

[116] Krishna Kanth Nakka, Xue Jiang, Dmitrii Usynin, and Xuebing Zhou. Pii jailbreaking in llms via activation steering reveals personal information leakage, 2025. Preprint.

[117] Kyle O'Brien, David Majercak, Xavier Fernandes, Richard Edgar, Blake Bullwinkel, Jingya Chen, Harsha Nori, Dean Carignan, Eric Horvitz, and Forough Poursabzi-Sangdeh. Steering language model refusal with sparse autoencoders. *arXiv preprint arXiv:2411.11296*, 2024.

[118] Hyunjong Ok, Jegwang Ryu, and Jaeho Lee. Decoding with limited teacher supervision requires understanding when to trust the teacher. *ArXiv*, abs/2406.18002, 2024.

[119] Deonna M Owens, Ryan A Rossi, Sungchul Kim, Tong Yu, Franck Dernoncourt, Xiang Chen, Ruiyi Zhang, Jiuxiang Gu, Hanieh Deilamsalehy, and Nedim Lipka. A multi-llm debiasing framework. *arXiv preprint arXiv:2409.13884*, 2024.

[120] Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre L. Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, Giulio Zizzo, Muhammad Zaid Hameed, Mark Purcell, Michael Desmond, Qian Pan, Inge Vejsbjerg, Elizabeth Daly, Michael Hind, Werner Geyer, Ambrish Rawat, Kush R. Varshney, and Prasanna Sattigeri. Granite guardian. *ArXiv*, abs/2412.07724, 2024.

[121] Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *AAAI Conference on Artificial Intelligence*, 2024.

[122] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*, 2023.

[123] Nicholas Pochinkov and Nandi Schoots. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*, 2024.

[124] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, 2023.

[125] Senmao Qi, Yifei Zou, Peng Li, Ziyi Lin, Xiuzhen Cheng, and Dongxiao Yu. Amplified vulnerabilities: Structured jailbreak attacks on llm-based multi-agent debate, 2025.

[126] Zhanyue Qin, Haochuan Wang, Zecheng Wang, Deyuan Liu, Cunhang Fan, Zhao Lv, Zhiying Tu, Dianhui Chu, and Dianbo Sui. Mitigating gender bias in code large language models via model editing. *arXiv preprint arXiv:2410.07820*, 2024.

[127] Ruizhong Qiu, Gaotang Li, Tianxin Wei, Jingrui He, and Hanghang Tong. Saffron-1: Towards an inference scaling paradigm for llm safety assurance. *arXiv preprint arXiv:2506.06444*, 2025.

[128] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

[129] Traian Rebedea, Razvan Laurentiu Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

[130] Debdeep Sanyal and Murari Mandal. Agents are all you need for llm unlearning. *arXiv preprint arXiv:2502.00406*, 2025.

[131] Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, Jorrit Kruthoff, Scott Goodfriend, Euan Ong, Alwin Peng, Raj Agarwal, Cem Anil, Amanda Askell, Nathan Bailey, Joe Benton, Emma Bluemke, Samuel R. Bowman, Eric Christiansen, Hoagy Cunningham, Andy Dau, Anjali Gopal, Rob Gilson, Logan Graham, Logan Howard, Nimit Kalra, Taesung Lee, Kevin Lin, Peter Lofgren, Francesco Mosconi, Clare O'Hara, Catherine Olsson, Linda Petrini, Samir Rajani, Nikhil Saxena, Alex Silverstein, Tanya Singh, Theodore Sumers, Leonard Tang, Kevin K. Troy, Constantin Weisser, Ruiqi Zhong, Giulio Zhou, Jan Leike, Jared Kaplan, and Ethan Perez. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025.

[132] Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. *arXiv preprint arXiv:2506.07022*, 2025.

[133] Luohe Shi, Yao Yao, Z. Li, Lefei Zhang, and Hai Zhao. Reference trustable decoding: A training-free augmentation paradigm for large language models. *ArXiv*, abs/2409.20181, 2024.

[134] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8364–8377, 2024.

[135] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. Fairrag: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12005, 2024.

[136] Chengyu Song, Linru Ma, Jianming Zheng, Jinzhi Liao, Hongyu Kuang, and Lin Yang. Audit-llm: Multi-agent collaboration for log-based insider threat detection, 2024.

[137] Xiaoxi Sun, Jinpeng Li, Yan Zhong, Dongyan Zhao, and Rui Yan. Towards detecting llms hallucination via markov chain-based multi-agent debate framework. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[138] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29904–29914, 2025.

[139] Manan Suri, Nishit Anand, and Amisha Bhaskar. Mitigating memorization in llms using activation steering, 2025. Preprint.

[140] Shota Takashiro, Takeshi Kojima, Andrew Gambardella, Qi Cao, Yusuke Iwasawa, and Yutaka Matsuo. Answer when needed, forget when not: Language models pretend to forget via in-context knowledge unlearning, 2025.

[141] Ziyi Tang, Ruilin Wang, Weixing Chen, Yongsen Zheng, Zechuan Chen, Yang Liu, Keze Wang, Tianshui Chen, and Liang Lin. Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms, 2025.

[142] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. Large language models post-training: Surveying techniques from alignment to reasoning. *arXiv preprint arXiv:2503.06072*, 2025.

[143] Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, and Zhikun Zhang. Inferdpt: Privacy-preserving inference for black-box large language models. *IEEE Transactions on Dependable and Secure Computing*, 2025.

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

[145] Bibek Upadhayay, Ph.D Vahid Behzadan, Minjia Wang, Pingping Lin, Siqi Cai, Shengnan An, Shengjie Ma, Zeqi Lin, Congrui Huang, Alexander Wei, Nika Haghtalab, Jacob Steinhardt, Jailbroken, Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Huai hsin Chi, V Quoc, Denny Le, Zhou, Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, Yisen Wang, Xiaofei Wen, Wenjie Wenxuan Zhou, Jacky Mo, Thinkguard, Zhen Xiang, Linzhi Zheng, Yanjie Li, Junyuan Hong, Qinbin Li, Han Xie, Jiawei Zhang, Zidi Xiong, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, and Karthik Narasimhan. X-guard: Multilingual guard agent for content moderation. *ArXiv*, abs/2504.08848, 2025.

[146] U.S. AI Safety Institute at NIST. Managing misuse risk for dual-use foundation models (nist ai 800-1, 2nd public draft). Technical report, NIST, 2025.

[147] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 2017.

[148] David Wan, Justin Chen, Elias Stengel-Eskin, and Mohit Bansal. MAMM-refine: A recipe for improving faithfulness in generation with multi-agent collaboration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9882–9901, 2025.

[149] Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. Dell: Generating reactions and explanations for llm-based misinformation detection. In *ACL (Findings)*, 2024.

[150] Yixin Wan and Kai-Wei Chang. White men lead, black women help? benchmarking and mitigating language agency social biases in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9082–9108, 2025.

[151] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

[152] Yixin Wan, Xingrun Chen, and Kai-Wei Chang. Which cultural lens do models adopt? on cultural positioning bias and agentic mitigation in llms, 2025. URL https://arxiv.org/abs/2509.21080.

[153] Chenxi Wang, Xiang Chen, Ningyu Zhang, Bo Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *ArXiv*, abs/2410.11779, 2024.

[154] Haoran Wang, Xiongxiao Xu, Baixiang Huang, and Kai Shu. Privacy-aware decoding: Mitigating privacy leakage of large language models in retrieval-augmented generation. 2025.

[155] Liwen Wang, Wenxuan Wang, Shuai Wang, Zongjie Li, Zhenlan Ji, Zongyi Lyu, Daoyuan Wu, and Shing-Chi Cheung. Ip leakage attacks targeting llm-based multi-agent systems, 2025.

[156] Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*, 2024.

[157] Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories, 2024. Also appears at TheWebConf (WWW) 2025.

[158] Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36:74530–74543, 2023.

[159] Xunguang Wang, Zhenlan Ji, Wenxuan Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. Sok: Evaluating jailbreak guardrails for large language models. *ArXiv*, abs/2506.10597, 2025.

[160] Yanbo Wang, Jiayi Ye, Siyuan Wu, Chujie Gao, Yue Huang, Xiuying Chen, Yue Zhao, and Xiangliang Zhang. TrustEval: A dynamic evaluation toolkit on trustworthiness of generative foundation models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 2025.

[161] Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *arXiv preprint arXiv:2502.19649*, 2025.

[162] Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

[163] Jiankun Wei, Abdulrahman Abdulrazzag, Tianchen Zhang, Adel Muursepp, and Gururaj Saileshwar. Privacy risks of speculative decoding in large language models. *ArXiv*, abs/2411.01076, 2024.

[164] Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023.

[165] Yixin Wu, Rui Wen, Michael Backes, Pascal Berrang, Mathias Humbert, Yun Shen, and Yang Zhang. Quantifying privacy risks of prompts in visual prompt learning. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 5841–5858, 2024.

[166] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. *ArXiv*, abs/2402.08983, 2024.

[167] Han Yan, Zheyuan Liu, and Meng Jiang. Dual-space smoothness for robust and balanced llm unlearning. *arXiv preprint arXiv:2509.23362*, 2025.

[168] Ruixin Yang, Dheeraj Rajagopal, Shirley Anugrah Hayati, Bin Hu, and Dongyeop Kang. Confidence calibration and rationalization for llms via multi-agent deliberation, 2024.

[169] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

[170] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pang, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. A survey on trustworthy llm agents: Threats and countermeasures. *arXiv preprint arXiv:2503.09648*, 2025.

[171] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, S Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In *International Conference on Learning Representations*, 2023.

[172] Fan Yuan, Chi Qin, Xiaogang Xu, and Piji Li. Helpd: Mitigating hallucination of lvlms by hierarchical feedback learning with vision-enhanced penalty decoding. *ArXiv*, abs/2409.20429, 2024.

[173] Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22484–22492, 2024.

[174] Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative ai content moderation based on gemma. *ArXiv*, abs/2407.21772, 2024.

[175] Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. Shieldgemma 2: Robust and tractable image content moderation. *ArXiv*, abs/2504.01081, 2025.

[176] Xinyi Zeng, Yuying Shang, Yutao Zhu, Jiawei Chen, and Yu Tian. Root defence strategies: Ensuring safety of llm at the decoding level. *ArXiv*, abs/2410.06809, 2024.

[177] Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. Autodefense: Multi-agent LLM defense against jailbreak attacks. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[178] Jiawen Zhang, Kejia Chen, Lipeng He, Jian Lou, Dan Li, Zunlei Feng, Mingli Song, Jian Liu, Kui Ren, and Xiaohu Yang. Activation approximations can incur safety vulnerabilities even in aligned llms: Comprehensive analysis and defense. *arXiv preprint arXiv:2502.00840*, 2025.

[179] Jingyu (Jack) Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *ArXiv*, abs/2410.08968, 2024.

[180] Zaibin Zhang, Yongting Zhang, Lijun Li, Hongzhi Gao, Lijun Wang, Huchuan Lu, Feng Zhao, Yu Qiao, and Jing Shao. PsySafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15202–15231, 2024.

[181] Zhexin Zhang, Junxiao Yang, Yida Lu, Pei Ke, Shiyao Cui, Chujie Zheng, Hongning Wang, and Minlie Huang. From theft to bomb-making: The ripple effect of unlearning in defending against jailbreak attacks. *arXiv preprint arXiv:2407.02855*, 2024.

[182] Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*, 2025.

[183] Zhengyue Zhao, Xiaoyun Zhang, Kaidi Xu, Xing Hu, Rui Zhang, Zidong Du, Qi Guo, and Yunji Chen. Adversarial contrastive decoding: Boosting safety alignment of large language models via opposite prompt optimization. *ArXiv*, abs/2406.16743, 2024.

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

[185] Jingnan Zheng, Xiangtian Ji, Yijun Lu, Chenhang Cui, Weixiang Zhao, Gelei Deng, Zhenkai Liang, An Zhang, and Tat-Seng Chua. Rsafe: Incentivizing proactive reasoning to build robust and adaptive llm safeguards. *ArXiv*, abs/2506.07736, 2025.

[186] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.

[187] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. Rose doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

[188] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, 2023.

[189] Andy Zhou. Compositional subspace representation fine-tuning for adaptive large language models. *arXiv preprint arXiv:2503.10617*, 2025.

[190] Jialong Zhou, Lichao Wang, and Xiao Yang. Guardian: Safeguarding llm multi-agent collaborations with temporal graph modeling, 2025.

[191] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, 2023.

[192] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.

[193] Yujia Zhou, Zheng Liu, Jiajie Jin, Jian-Yun Nie, and Zhicheng Dou. Metacognitive retrieval-augmented large language models. In *Proceedings of the ACM Web Conference 2024*, pages 1453–1463, 2024.

[194] Yujun Zhou, Yufei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. Defending jailbreak prompts via in-context adversarial game. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20084–20105, 2024.

[195] Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*, 2024.

[196] Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.

[197] Yifan Zhu, Chao Zhang, Xin Shi, Xueqiao Zhang, Yi Yang, and Yawei Luo. Master: Multi-agent security through exploration of roles and topological structures – a comprehensive framework, 2025.

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

[199] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *ArXiv*, abs/2307.15043, 2023.

[200] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, 2025.