Guidance Large Language Models at Test Time: A Unified Review of LLM-Training-Free Methods

Anonymous ACL submission

Abstract

Adapting Large Language Models (LLMs) to dynamic constraints typically requires expensive fine-tuning. While training-free test-time guidance offers a flexible alternative, the literature remains fragmented across isolated subfields. This paper presents a unified review of LLM-training-free guidance, systematizing methods that steer behavior without parameter updates. We propose a taxonomy based on the inference lifecycle, categorizing interventions into Input-Space, Latent-Space, Decoding-Space, and Output-Space guidance. Furthermore, we analyze critical trade-offs regarding model accessibility, computational cost, and control granularity. Finally, we discuss emerging frontiers, highlighting the convergence of control mechanisms towards unified architectures and the shift toward rigorous, interpretability-driven steering.

1 Introduction

Large Language Models (LLMs) have evolved into general-purpose foundations capable of solving diverse tasks (OpenAI et al., 2024; Touvron et al., 2023; Guo et al., 2025). To adapt these broad models to specific domains or safety policies, the standard approach is Post-Training, such as Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Rafailov et al., 2023). While effective, relying on parameter updates creates three significant bottlenecks in real-world deployment. First, the process is computationally expensive, as iteratively re-training models for every new requirement consumes massive resources (Hu et al., 2022; Ning et al., 2025). Second, it involves significant risks, specifically "catastrophic forgetting" (Li et al., 2024a), where optimizing for a narrow task degrades the model's general capabilities (Harmon et al., 2025). Finally, adaptation is often impossible in commercial settings where models are deployed as black boxes, blocking user access to weight modification.

Fortunately, parameter updates are not the sole mechanism for adaptation. A parallel and flexible paradigm involves steering the inference process itself while keeping the base model frozen. This approach is pluggable, reversible, and highly efficient, allowing practitioners to modulate behavior on-the-fly. In practice, interventions can occur at various levels: from injecting external knowledge via Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Fan et al., 2024) to adjusting internal neuron activations for surgical style control (Turner et al., 2024). Other techniques involve filtering token probabilities to enforce strict constraints (Lu et al., 2021) or employing iterative search algorithms for complex reasoning (Yao et al., 2023a; Besta et al., 2024). Notably, recent evidence suggests that simple, well-structured prompting can sometimes outperform complex fine-tuning, challenging the assumption that heavy engineering is always necessary (Zhou et al., 2023a; Schulhoff et al., 2025).

Despite this momentum, the research landscape remains fragmented. Broad surveys typically focus on general model capabilities (Zhao et al., 2025a) or post-training scaling (Lai et al., 2025), while reviews on prompting (Schulhoff et al., 2025), retrieval (Fan et al., 2024), and interpretability (Rai et al., 2025) are usually studied in isolation. Although recent works explore test-time scaling (Zhang et al., 2025b), cooperation (Huang et al., 2025a), and alignment (Pan et al., 2025), they prioritize specific objectives rather than offering a holistic view of the control mechanism(see Table 1 in Appendix for a detailed scope comparison). There lacks a mechanism-centric unified framework that connects the entire inference lifecycle, leaving methods that operate on similar principles treated as unrelated simply because they intervene at different stages.

This survey aims to fill this gap by presenting a unified Test-Time Lifecycle Taxonomy for LLM-Training-Free Guidance. We treat test-time guidance as a holistic control problem, organizing methods based on *where* the intervention occurs in the generation process. This perspective allows us to compare disparate techniques—from simple system prompts to surgical activation guidance—on a single map. It also reveals common trade-offs regarding access assumptions (Black-box vs. White-box), control granularity, and test-time latency. We use the term *LLM-Training-Free* to mean that the target LLM weights are not updated, while allowing the use of pluggable auxiliary components (Cunningham et al., 2023).

This survey makes three key contributions:

- **Lifecycle Taxonomy.** We synthesize fragmented methods into a unified framework spanning Input, Latent, Decoding, and Output spaces (Sec.2).
- **Analysis.** We analyze the capabilities of each paradigm regarding model access, inference cost, and steering precision (Sec.3).
- **Future Frontiers.** We identify the emerging trend of "full-stack" architectures and interpretability-driven control mechanisms (Sec.4.2).

2 Problem Formulation and Taxonomy

This section formalizes the concept of *LLM-Training-Free Test-Time Guidance* and introduces a taxonomy organized by the specific stage within the inference lifecycle where an intervention occurs.

2.1 Definition and Problem Formulation

Consider a target Large Language Model (LLM) parameterized by θ, which defines an autoregressive conditional distribution over an output sequence $y = (y_1, \ldots, y_T)$ given an input context x:

$$p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid x, y_{<t}) \quad (1)$$

Test-Time Guidance. We define *test-time guidance* as the introduction of a guidance operator $\mathcal{G}$ that transforms the original generative process into a guided distribution:

$$p_{\theta,\mathcal{G}}(y \mid x, g) = \mathcal{G}(p_\theta(\cdot \mid x), g) \quad (2)$$

where g denotes the *guidance signal*. This signal can manifest as textual instructions (Wei et al., 2022), a latent direction vector (Turner et al., 2024; Zou et al., 2023), a decoding penalty (Li et al., 2023b), or a verifier scoring function (Yao et al., 2023a). Equation 2 is deliberately broad to encompass hybrid systems that may combine multiple signals (e.g., a safety system prompt combined with logit-level constraints).

LLM-Training-Free Constraint. The term *LLM-Training-Free* imposes a strict constraint on the *target* model parameters: during guidance, θ must remain frozen ($\Delta\theta = \mathbf{0}$). This explicitly excludes methods that update the base model, such as full fine-tuning (Devlin et al., 2019) or LoRA (Hu et al., 2022). Crucially, this definition *does not* forbid the use of external auxiliary modules, even if those modules are trained, as long as they are structurally separate from the LLM. Representative examples include dense retrievers (Karpukhin et al., 2020), sparse autoencoders (Huben et al., 2024), and reward models (Stiennon et al., 2022). This modularity is a key practical advantage, allowing us to upgrade control mechanisms without the high computational cost and stability risks associated with re-training the base model.

2.2 Taxonomy: A Lifecycle View

We classify guidance methods based on the principal interface through which $\mathcal{G}$ intervenes. As illustrated in Figure 1, this taxonomy maps methods to four distinct stages of the inference lifecycle: the input context, the internal representation, the decoding distribution, and the output space.

Input-Space Guidance. Input-space guidance transforms the raw input x into a modified context x' using the guidance signal g:

$$x' = f(x; g), \quad p_{\theta,\mathcal{G}}(y \mid x, g) = p_\theta(y \mid x') \quad (3)$$

Here, the guidance signal g typically manifests as demonstrations for In-Context Learning (ICL) (Brown et al., 2020; Min et al., 2022), external knowledge injection via Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Ram et al., 2023), or system-level instructions (commonly known as system prompts) (Ouyang et al., 2022; Touvron et al., 2023).

Latent-Space Guidance. Latent-space guidance intervenes directly on the internal representations computed during the forward pass (Zou et al., 2023;

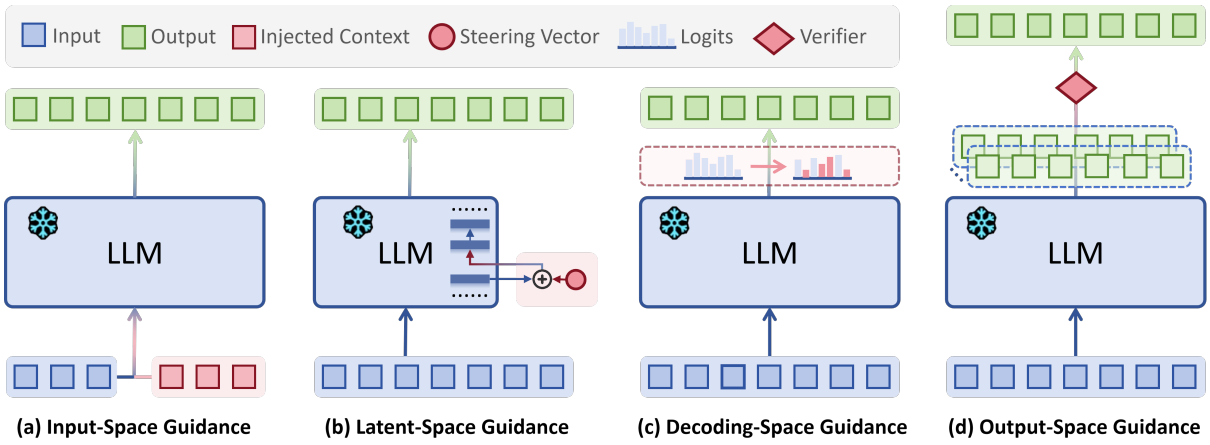

Figure 1: **Overview of the four paradigms for LLM-Training-Free Test-Time Guidance.** The snowflake denotes that the base model parameters remain frozen, while red elements highlight the intervention of guidance signals at distinct stages of the inference lifecycle.

Li et al., 2023a). Let h_t^ℓ denote the hidden state at layer ℓ and time step t. A latent intervention applies a transformation:

$$\tilde{h}_t^\ell = \mathcal{T}_\ell(h_t^\ell; g) \tag{4}$$

where subsequent layers consume the modified $\tilde{h}_t^\ell$. The guidance signal g typically encodes a direction in the semantic space (Turner et al., 2024; Zou et al., 2023) or a sparse feature clamp (Huben et al., 2024; Yan et al., 2024). Unlike input guidance, which operates on discrete tokens, latent guidance manipulates the continuous semantic manifold of the model, offering finer granularity for attributes such as sentiment (Subramani et al., 2022) or hallucination mitigation (Li et al., 2023a).

Decoding-Space Guidance. Decoding-space guidance intervenes directly at the token selection interface, reshaping the local probability landscape without updating parameters. Let $z_t \in R^{|V|}$ denote the raw logits at time step t. We formalize the intervention as a logit transformation function r applied prior to normalization:

$$\tilde{z}_t = r(z_t, x, y_{<t}; g) \tag{5}$$

$$p_{\theta,\mathcal{G}}(y_t \mid x, y_{<t}) = \text{softmax}(\tilde{z}_t) \tag{6}$$

The operator r can enforce hard lexical constraints via masking (Lu et al., 2021), or implement soft contrastive adjustments (Li et al., 2023b; Chuang et al., 2024) where logits from an "amateur" model or early layers are subtracted from z_t. This approach directly modulates the output distribution, providing robust control over specific attributes (Yang and Klein, 2021) and formatting, though it requires balancing steering strength against potential coherence degradation.

Output-Space Guidance. Output-space guidance shifts the locus of intervention from the step-by-step generation process to the evaluation of complete sequences. Rather than shaping local token probabilities, the system generates a set of candidate hypotheses and selects the optimal output y^*:

$$y^* = \arg\max_{y \in \mathcal{C}} R(y; x, g) \tag{7}$$

Here, $\mathcal{C}$ represents candidates sampled from the base distribution $p_\theta(\cdot|x)$, while R acts as a post-hoc filter or a verifier (Cobbe et al., 2021) or search heuristic (Yao et al., 2023a). This paradigm treats the frozen LLM purely as a proposal distribution, delegating alignment entirely to the selection process.

3 Guidance Methodologies

3.1 Input-Space Guidance

Input-space guidance steers black-box models by systematically transforming the query x rather than internal parameters. As summarized in Figure 2, we categorize these structured frameworks into *Contextual Augmentation*, *Prompt Design and Optimization*, and *Alignment and Safety Prompting*.

3.1.1 Contextual Augmentation

This paradigm bridges the gap between parametric memory and query needs by injecting external information. Retrieval-Augmented Generation (RAG) has evolved from static passage injection (Ram et al., 2023; Shi et al., 2024b) to active,

Contextual Augmentation

IC-RAG (Ram et al., 2023), REPLUG (Shi et al., 2024b), CoK (Li et al., 2024b), RQ-RAG (Chan et al., 2024), GMR (Lee et al., 2022), RAG-Fusion (Rackauckas, 2024), KRA-GEN (Matsumoto et al., 2024), SimRAG (Xu et al., 2025), SEER (Zhao et al., 2024a), L-RAG (Lin et al., 2025a), HopRAG (Liu et al., 2025a), KERAG (Sun et al., 2025c), TC-RAG (Jiang et al., 2025b), HyKGE (Jiang et al., 2025c), UniRAG (Li et al., 2025e), Om-niRAGMed (Chen et al., 2025c), MolRAG (Xian et al., 2025), EventRAG (Yang et al., 2025d), KiRAG (Fang et al., 2025a), FaithfulRAG (Zhang et al., 2025a), Dialogue-RAG (Li et al., 2025d), SGIC (Chen et al., 2025a), MedGraphRAG (Wu et al., 2025a), Astute RAG (Wang et al., 2025a), DualRAG (Cheng et al., 2025), T-GRAG (Li et al., 2025b), SafeDriveRAG (Ye et al., 2025), RecipeRAG (Yang et al., 2025c), HM-RAG (Liu et al., 2025c), DynamicRAG (Sun et al., 2025b), What-makes-ICL (Liu et al., 2022), Lost-in-Middle (Liu et al., 2024a), L2R-ICL (Wang et al., 2024a), CED-Select (Iter et al., 2023), RGER (Lin et al., 2025b), GENICL (Zhang et al., 2025e), MLSM/TTF (Liu et al., 2025b), UniICL (Gao et al., 2025), FLARE (Jiang et al., 2023b), CoVe (Wang et al., 2024c), ToG2(Ma et al., 2025)

Prompt Design and Optimization

Chain-of-Thought (Wei et al., 2022), Zero-shot CoT (Kojima et al., 2022), Self-Consistency (Wang et al., 2023), Tab-CoT (Ziqi and Lu, 2023), Auto-CoT (Zhang et al., 2023), Logical CoT (Zhao et al., 2024b), CDW-CoT (Fang et al., 2025b), RankCoT (Wu et al., 2025c), PCoT (Modzelewski et al., 2025), SemCoT (He et al., 2025a), APE (Zhou et al., 2023b), OPRO (Yang et al., 2024), BPO (Cheng et al., 2024a), RLPrompt (Deng et al., 2022), PromptAgent (Wang et al., 2024b), PEARL-plan (Sun et al., 2024b), PromptBreeder (Fernando et al., 2024), Structure-Guided (Cheng et al., 2024b), AutoPrompt (Shin et al., 2020), Automate-CoT (Shum et al., 2023)

Alignment and Safety Prompting

Self-Reminders (Xie et al., 2023), URIAL (Lin et al., 2024), LLaMA-2 cfg. (Touvron et al., 2023), ICAG (Zhou et al., 2024b), SmoothLLM (Robey et al., 2025), ROSE (Zhong et al., 2024) , Intention Analysis (Zhang et al., 2025d), CIP (Hahm et al., 2025), LLM-Self-Defense (Phute et al., 2024), SelfDefenD (Wang et al., 2025f), P-Aligner (Song et al., 2025b), ICDPO (Song et al., 2025a)

Figure 2: Input-Space guidance methods.

structure-aware inference. Recent frameworks interleave retrieval with reasoning steps (Li et al., 2024b; Chan et al., 2024; Rackauckas, 2024) or leverage graph-based and domain-specific structures—such as knowledge graphs (Liu et al., 2025a; Wu et al., 2025a), molecular fingerprints (Xian et al., 2025), and interaction histories (Li et al., 2025d)—to handle complex queries. Active strategies further empower models to decide *when* to retrieve (Jiang et al., 2023b; Yao et al., 2023b), optimizing the information flow dynamically.

Parallelly, In-Context Learning (ICL) optimization focuses on the precise selection and compression of demonstrations. Instead of simple similarity matching, modern approaches align exemplars with task-specific reasoning patterns (Iter et al., 2023; Lin et al., 2025b; Zhang et al., 2025e) or compress extensive contexts into efficient "memory tokens" to reduce computational overhead (Gao et al., 2025; Liu et al., 2025b).

3.1.2 Prompt Design and Optimization

Treating the prompt as a programmable interface, this stream optimizes the instruction format to elicit specific capabilities. Structured Reasoning methods extend standard Chain-of-Thought (Wei et al., 2022) by imposing strict logical, tabular, or semantic constraints (Ziqi and Lu, 2023; Zhao et al., 2024b; He et al., 2025a) and ranking reasoning paths (Wu et al., 2025c) to minimize hallucinations.

To eliminate manual trial-and-error, Automated Prompt Optimization formulates instruction design as a search problem. Black-box optimization methods utilize task feedback to iteratively refine prompts (Yang et al., 2024; Zhou et al., 2023b; Cheng et al., 2024a), while advanced planning algorithms employ Monte Carlo Tree Search (Wang et al., 2024b) or evolutionary strategies (Fernando et al., 2024; Cheng et al., 2024b) to navigate the prompt space, effectively turning prompt engineering into an algorithmic search process.

3.1.3 Alignment and Safety Prompting

Input-space guidance also serves as a critical, training-free defense layer. While early work relied on static system prompts to define behavioral boundaries (Lin et al., 2024; Touvron et al., 2023; Xie et al., 2023), 2025 methodologies have shifted towards Active Instance-Level Alignment. Approaches like P-Aligner and ICDPO dynamically rewrite user queries into principled forms based on alignment preferences (Song et al., 2025b,a). Furthermore, Agentic Defense mechanisms decouple safety evaluation from response generation; by employing shadow models or causal reasoning agents to screen inputs and monitor potential risks (Phute et al., 2024; Wang et al., 2025f; Hahm et al., 2025; Zhou et al., 2024b; Zhang et al., 2025d), these systems provide robust protection against jailbreaks and adversarial attacks without modifying model weights.

3.2 Latent-Space Guidance

Latent-space guidance intervenes on the internal representations of a frozen LLM during inference. Formally, a guidance operator transforms hidden states h_t^ℓ into $\tilde{h}_t^\ell$ via a transformation $\mathcal{T}_\ell(h_t^\ell; g)$ driven by a guidance signal g. Unlike input-space methods that manipulate discrete tokens, these approaches operate on continuous activation vectors, utilizing probes, autoencoders, or geometric projections to steer model behavior. We categorize these methodologies based on their geometric granularity and dynamic nature as follows.

Global Directional Arithmetic. The most foundational approach treats guidance as a global shift along learned directions in the residual stream, typically formalized as an additive update $\tilde{h}_t^\ell = h_t^\ell + \alpha v$. PPLM (Dathathri et al., 2020) pioneered this paradigm by utilizing test-time gradient updates to shift hidden states toward target attributes.

Activation Addition (Turner et al., 2024) and Representation Engineering (RepE) (Zou et al., 2023) extract these steering vectors by contrasting activations from paired positive and negative prompts, successfully modulating high-level traits like helpfulness, honesty, or political framing. Inference-Time Intervention (ITI) (Li et al., 2023a) derives truthfulness directions from linear probes trained on labeled datasets, while In-Context Vectors (ICV) (Liu et al., 2024b) and Function Vectors (Todd et al., 2024) distill task-specific directions directly from few-shot demonstrations. Recent empirical studies focus on optimizing the application of these vectors, analyzing layer sensitivity and scaling laws to maximize control efficacy while minimizing perplexity degradation (Stoehr et al., 2024; Stolfo et al., 2025; Liu et al., 2025d).

Decomposition and Fine-Grained Control. Moving beyond raw residual directions, this paradigm decomposes representations into an interpretable basis before intervention. Sparse Autoencoders (SAEs) have emerged as a powerful tool to disentangle polysemantic neurons into monosemantic features; steering in this context implies identifying features correlated with specific behaviors and clamping their coefficients (Huben et al., 2024; Soo et al., 2025; Yan et al., 2024). At the neuron level, methods identify "target atoms" or clusters of neurons responsible for hallucinations or toxic content, selectively dampening them to mitigate adverse behaviors while preserving general reasoning capabilities (Zhang et al., 2025f; Huang et al., 2025c; Han et al., 2025). Work in 2025 further refines this by mapping safety-critical concepts to sparse bases to decouple refusal mechanisms from reasoning logic (Wang et al., 2025c; Valentino et al., 2025; Sakarvadia et al., 2025).

Geometric Constraints and Projection. To minimize the side effects of steering, this category employs linear operators that respect the geometric structure of the representation space. Instead of simple addition, methods like AlphaSteer (Sheng et al., 2025) and ASGuard (Park et al., 2025) compute steering vectors within the null-space of task-relevant representations, ensuring that safety edits remain orthogonal to core competencies. Projection-based steering (Postmus and Abreu, 2024) and probe-free low-rank interventions (Jiang et al., 2025a; Oozeer et al., 2025) parameterize guidance as a low-rank update $\tilde{h} = h + Wh$, learning a structured transformation that aligns activations with a target distribution rather than a single direction (Rodriguez et al., 2024). These methods emphasize "surgical" interventions that mathematically constrain the edit distance in task-critical subspaces.

Latent-Space Guidance

Latent steering vectors (Subramani et al., 2022), IDANI (Antverg et al., 2022), PPLM (Dathathri et al., 2020), ITI (Inference-Time Intervention) (Li et al., 2023a), Activation Engineering (Turner et al., 2024), Representation Engineering (RepE) (Zou et al., 2023), (Huben et al., 2024), (Yan et al., 2024), EAST (Entropic Activation Steering) (Rahn et al., 2024), NL-ITI (Hoscilowicz et al., 2024), (Fatahi Bayat et al., 2024), In-context Vectors (ICV) (Liu et al., 2024b), Activation scaling (Stoehr et al., 2024), (Postmus and Abreu, 2024), Steering Target Atoms (STA) (Wang et al., 2025c), SCANS (Cao et al., 2025), FLORAIN (Jiang et al., 2025a), SADI (Wang et al., 2025e), DSAS (Do et al., 2025), (Valentino et al., 2025), (Sakarvadia et al., 2025), (Zhao et al., 2025b), (Oozeer et al., 2025), (Cyberey and Evans, 2025), AlphaSteer (Sheng et al., 2025), (Jorgensen et al., 2023), (van der Weij et al., 2024), (Todd et al., 2024), (Konen et al., 2024), (Stolfo et al., 2025), (Lee et al., 2025a), (Wang et al., 2025d), (Hegazy et al., 2025), (Rodriguez et al., 2024), (Zhang et al., 2025f), (Park et al., 2025), (Wu et al., 2025b), (Sun et al., 2025a), (Soo et al., 2025), (Karnik and Bansal, 2025), (Huang et al., 2025c), (Han et al., 2025), (Liu et al., 2025d), (Scalena et al., 2024)

Figure 3: Latent-space guidance methods.

Decoding-Space Guidance

GeDi (Krause et al., 2021), Contrastive Decoding (Li et al., 2023b), DoLa (Chuang et al., 2024), NeuroLogic Decoding (Lu et al., 2021), FUDGE (Yang and Klein, 2021), DExperts (Liu et al., 2021), Grammar-Constrained Decoding (Raspanti et al., 2025), ROSE (Zhong et al., 2024), Instructive Decoding (Alajrami et al., 2025), MOD (Shi et al., 2024a), RMOD (Son et al., 2025), PAD (Chen et al., 2025b), DeAL (Huang et al., 2025b), DeRa (Liu et al., 2024c), SAFEINFER (Banerjee et al., 2025), IBD (Zhu et al., 2024), Nudging (Fei et al., 2025), CDT (Yang et al., 2025a), UACD (Lee et al., 2025b), SLED (Zhang et al., 2024), RGD (Mañas et al., 2025), PRGD (Nguyen et al., 2025), ACD (Zhang et al., 2025c), Drift (Kim et al., 2025), RSD (Liao et al., 2025), Air-Decoding (Zhong et al., 2023), CARDS (Li et al., 2025a), CARE (Hu et al., 2025b), CDT (Yang et al., 2025b), MCA (Fu et al., 2025).

Figure 4: Decoding-space guidance methods.

Context-Adaptive Dynamics. Fixed interventions may lead to over-control. Adaptive methods resolve this by modulating the intervention strength α or direction v based on the current context. Nonlinear ITI (Li et al., 2023a) and entropy-aware steering (Hoscilowicz et al., 2024) scale interventions based on model uncertainty, applying corrections only when hallucination risk is high. Semantics-adaptive controllers (Wang et al., 2025e,d) dynamically adjust activation scales according to the semantic topic of the input. In safety domains, methods like SCANS (Cao et al., 2025) and preemptive detectors (Karnik and Bansal, 2025) employ lightweight gating mechanisms to trigger steering only when harmful intent is detected. Extensions to multimodal settings, such as AutoSteer (Wu et al., 2025b) and Layer-Navigator (Sun et al., 2025a), dynamically route guidance across layers depending on joint vision-language activation patterns, ensuring context-sensitive safety alignment.

3.3 Decoding-Space Guidance

Decoding-space guidance intervenes on the logits z_t at each generation step, modifying token

probabilities via a guidance operator r before sampling. This enables fine-grained steering of style, safety, and task behavior without parameter updates. As summarized in Figure 4, we categorize these methodologies into *Score-based Reweighting* and *Constraint-based Filtering*.

Score-based Reweighting. Most decoding-time methods retain the full vocabulary but adjust the relative probability of tokens to favor desirable behaviors. GeDi (Krause et al., 2021) and DExperts (Liu et al., 2021) achieved this by reweighting predictions using logits from external auxiliary models (experts) or anti-experts. Contrastive strategies leverage the model's own capabilities to "self-correct" logits: methods like Contrastive Decoding (Li et al., 2023b) and DoLa (Chuang et al., 2024) subtract the logits of a "weak" model from a "strong" one to amplify factual signals. Newer variants such as UACD (Lee et al., 2025b) and SLED (Zhang et al., 2024) introduce uncertainty and stability metrics to modulate this contrastive penalty, avoiding noise when the model is unconfident. In safety alignment, ROSE (Zhong et al., 2024), Instructive Decoding (Alajrami et al., 2025), and ACD (Zhang et al., 2025c) down-weight tokens favored by unsafe or noisy prompts while boosting those aligned with clean instructions. A second cluster fuses base logits with scores from external auxiliary models or reward functions. Frameworks like MOD (Shi et al., 2024a) and DeAL (Huang et al., 2025b) combine task utility scores with base probabilities, while PAD (Chen et al., 2025b) and Drift (Kim et al., 2025) inject user-specific preference vectors for personalized alignment. Similarly, methods such as SAFEINFER (Banerjee et al., 2025) and CDT (Yang et al., 2025b) fuse risk estimation or truthfulness comparator scores to dynamically suppress toxic or hallucinatory continuations.

Constraint-based Filtering. A complementary paradigm treats decoding as a constrained search problem, using the guidance operator to strictly rule out invalid tokens rather than merely reshaping their scores. Classical methods employ symbolic masks to enforce validity: NeuroLogic Decoding (Lu et al., 2021) applies predicate logic constraints, while Grammar-Constrained Decoding (Raspanti et al., 2025) uses formal grammars to ensure syntactically valid code generation. Extending this to dynamic trajectory control, recent approaches explicitly prune the search space based on intermediate quality assessments. CARE (Hu et al., 2025b) introduces a "rollback and introspection" mechanism that detects unsafe trajectories mid-generation and reverts to a safe state. Similarly, CARDS (Li et al., 2025a) and Reward-Guided Speculative Decoding (RSD) (Liao et al., 2025) implement cascade sampling or verifiers to filter out low-quality speculative candidates before they are committed, effectively pruning undesirable paths to guarantee alignment constraints.

Output-Space Guidance

ToT (Yao et al., 2023a), GoT (Besta et al., 2024), RAP (Hao et al., 2023), Multi-Agent Debate (Zhou et al., 2024a), Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023), CRITIC (Gou et al., 2024), Training Verifiers (Cobbe et al., 2021), SelfCheckGPT (Manakul et al., 2023), MBR Decoding (Suzgun et al., 2023), Universal Self-Consistency (Chen et al., 2024), Speculative Rejection (Sun et al., 2024a),((Ren et al., 2023)), COOPERATE(Feng et al., 2024), LLM-Blender (Jiang et al., 2023a), Selection-Inference (Creswell et al., 2023), Mutual Reasoning (Qi et al., 2025a)

Figure 5: Output-space guidance methods.

3.4 ◆ Output-Space Guidance

Output-space guidance steers generation at the response level via post-hoc scoring or structural search, enabling complex reasoning without parameter updates. As summarized in Figure 5, we categorize these methods into *Search and Planning*, *Iterative Refinement*, *Candidate Selection and Reranking*, and *Multi-Model Collaboration*.

Search and Planning. These methods structure generation as a search process over a space of reasoning steps. Tree of Thoughts (ToT) (Yao et al., 2023a) allows the model to explore multiple branches and backtrack when necessary. Graph of Thoughts (GoT) (Besta et al., 2024) extends this to arbitrary graphs, enabling the aggregation of information from different thoughts. RAP (Hao et al., 2023) and LATS (Zhou et al., 2024a) treat the LLM as a world model within a Monte Carlo Tree Search framework, simulating future outcomes to plan the optimal reasoning path.

Iterative Refinement. Instead of searching multiple branches, refinement methods optimize a single response through feedback loops. Self-Refine (Madaan et al., 2023) generates an initial draft and iteratively prompts the model to critique and improve its own output. Reflexion (Shinn et al., 2023) introduces "verbal reinforcement learning," where the agent reflects on past failures to update its memory for future trials. CRITIC (Gou et al., 2024) enhances this cycle by using external tools to verify outputs and guide revisions based on execution

feedback.

Candidate Selection and Reranking. This paradigm follows a "generate-then-select" workflow: sample multiple candidates and select the best one. Verifier-based methods (Cobbe et al., 2021) and SelfCheckGPT (Manakul et al., 2023) score candidates based on correctness or hallucination probability. Universal Self-Consistency (Chen et al., 2024) and MBR Decoding (Suzgun et al., 2023) select the consensus output that minimizes risk or maximizes consistency across diverse samples. To improve efficiency, Speculative Rejection (Sun et al., 2024a) and Self-Evaluation (Ren et al., 2023) use reward models or confidence scores to prune low-quality candidates early in the process.

Multi-Model Collaboration. Guidance can also emerge from the interaction of diverse models. LLM-Blender (Jiang et al., 2023a) employs a ranking and fusion module to synthesize the best components from multiple model outputs. Selection-Inference (Creswell et al., 2023) decomposes reasoning into modular selection and inference steps handled by specialized prompts. Recent collaborative frameworks like Mutual Reasoning (Qi et al., 2025a) and COOPERATE (Feng et al., 2024) demonstrate that diverse models can correct each other's errors through voting or abstention mechanisms, leveraging the wisdom of the crowd without training.

4 Discussion

We synthesize these paradigms into a comparative framework to highlight trade-offs in access, cost, and granularity, followed by an analysis of critical deployment challenges.

4.1 Comparative Analysis

The Access Divide: Black-box vs. White-box. The most immediate constraint in real-world deployment is model access. Input- and output-space methodologies treat the LLM purely as an API endpoint, making them the only viable options for proprietary models like GPT-5 or Gemini. Techniques such as Contextual Augmentation (Ram et al., 2023) and Tree of Thoughts (Yao et al., 2023a) scale naturally with stronger base models without requiring internal visibility. In contrast, latent- and decoding-space guidance necessitate access to activation vectors or vocabulary logits. While this restricts their application to open-weights models, it unlocks a level of "surgical" precision impossible via prompting alone. For instance, Activation Engineering (Turner et al., 2024) and Inference-Time Intervention (ITI) (Li et al., 2023a) can suppress specific hallucination patterns or safety risks by directly dampening the corresponding internal features, bypassing the model's tendency to rationalize instructions in the prompt.

The Precision-Coherence Trade-off. Does finer control always lead to better outcomes? Evidence suggests a delicate balance between steering strength and semantic coherence. Decoding-space methods like NeuroLogic (Lu et al., 2021) and Contrastive Decoding (Li et al., 2023b) enforce strict token-level constraints, effectively pruning the probability mass of undesirable outputs. However, aggressively reshaping the output distribution can force the model off its natural "manifold," leading to disjointed or ungrammatical text. Similarly, latent interventions face the challenge of "manifold distortion": as noted in recent studies on activation scaling (Stolfo et al., 2025; Liu et al., 2025d), pushing internal states too far along a steering vector often causes a spike in perplexity and a degradation of general capabilities. This contrasts with input-space methods, which, by operating through natural language, generally preserve fluency but often fail to enforce negative constraints.

Inference-Time Compute as the New Scaling Law. A parallel trend connects these paradigms through the lens of compute allocation. While latent and decoding methods typically incur negligible overhead—often just a single vector addition or logit subtraction per step—output-space guidance represents a fundamental shift towards "test-time scaling." By generating multiple candidates or searching through reasoning trees, methods like ToT (Yao et al., 2023a) and RAP (Hao et al., 2023) effectively trade inference latency for intelligence, enabling System-2 reasoning. This mirrors the observation that difficult reasoning tasks benefit more from verifying multiple paths than from optimizing a single forward pass. Thus, the choice of method often reduces to a resource decision: is it more efficient to "patch" a model's intuition via latent steering, or to spend compute on explicit search via output guidance?

Convergence of Control Spaces. Although categorized into distinct spaces, recent advancements indicate a blurring of boundaries where effective systems hybridize these approaches. Contrastive

Decoding (Li et al., 2023b), nominally a decoding method, relies heavily on engineered "negative" prompts (input-space) to derive its steering signal (Zhong et al., 2024). Similarly, Latent-space vectors are often distilled from In-Context Learning examples (Liu et al., 2024b), effectively projecting input demonstrations into the residual stream. Most notably, safety frameworks like CARE (Hu et al., 2025b) now integrate decoding-time monitoring with output-space rollback mechanisms. This convergence suggests that future guidance will likely be "full-stack": leveraging prompts for intent, latent/decoding methods for constraints, and output search for verification.

4.2 Challenges and Future Directions

While training-free guidance offers flexibility, it faces critical hurdles in efficiency and reliability. We highlight four fundamental contradictions and the emerging frontiers promising to resolve them.

The Alignment-Capability Dilemma and Glass-Box Steering. A persistent challenge in current interventions is the trade-off between safety and capability, often resulting in an alignment tax where steering a model towards safety degrades its general reasoning. This phenomenon largely arises because standard activation steering operates on entangled, polysemantic representations, where a single direction encodes multiple distinct concepts. Consequently, blindly adding a steering vector often induces manifold distortion, leading to spikes in perplexity or over-refusal. To resolve this, the field is transitioning from coarse-grained direction arithmetic to interpretability-driven steering. By leveraging Sparse Autoencoders to decompose dense activations into overcomplete, monosemantic features (Yan et al., 2024; Cunningham et al., 2023), future methods will move towards precise feature clamping. This technique allows for the surgical suppression of specific behaviors, such as deception or bias, without disrupting the broader semantic manifold, effectively enabling control with minimal side effects (He et al., 2025b).

Inference Efficiency and Unified Architectures. Guidance shifts the computational burden from training to inference, where heavy retrieval or iterative search mechanisms can prohibit real-time deployment. To address this, future systems must evolve into unified "Full-Stack" architectures that employ hierarchical routing to optimize the computational budget per query, rather than applying a uniform processing chain. A meta-controller assesses input complexity, directing simple prompts to lightweight latent interventions (System-1) while reserving expensive tree-search algorithms for high-stakes reasoning tasks (System-2). Recent breakthroughs in scaling test-time compute (Snell et al., 2024) and reasoning-star strategies (Qi et al., 2025b) demonstrate that such adaptive allocation can simulate the performance gains of reinforcement learning without parameter updates, establishing inference-as-alignment as a resource-efficient paradigm.

Reliability Crisis and Theoretical Guarantees. Real-world applications require satisfying conflicting constraints simultaneously (e.g., conciseness vs. detailed reasoning), often leading to destructive interference in linear steering combinations. Moving beyond empirical tuning requires establishing a *"Control Theory for LLMs."* Emerging research treating generation as a Markov Decision Process (MDP) suggests that verifiable safety properties may be achievable through constrained decoding policies (Kumar et al., 2025) and Pareto-optimal multi-objective optimization (Son et al., 2025), ensuring convergence to safe states.

The Generalization Gap and Multimodal Steering. As LLMs evolve into Multimodal Large Language Models, text-only guidance is becoming insufficient. Visual inputs often bypass text-based safety guardrails, leading to "visual jailbreaks" where adversarial images trigger harmful outputs despite safe textual prompts. The frontier lies in developing cross-modal interfaces that bridge this gap by establishing a shared control space. Emerging research focuses on using textual prompts to directly modulate visual encoders (Wu et al., 2025b) or intervening on joint vision-language representations (Wang et al., 2025b). This points toward a modality-agnostic control layer capable of aligning video, audio, and text generation simultaneously.

5 Conclusion

This survey systematizes training-free guidance through a mechanism-centric Inference Lifecycle Taxonomy. By unifying interventions across input, latent, decoding, and output spaces, we highlight key trade-offs in accessibility, cost, and precision. We envision a future shift toward "full-stack" architectures that dynamically bridge static models with evolving user needs.

Limitations

This survey establishes a structural foundation for training-free guidance, yet we acknowledge the gap between our theoretical taxonomy and engineering reality. Our mechanism-centric classification inevitably abstracts over complex implementation nuances—such as inference latency trade-offs and memory optimization—that are critical in deployment but distinct from conceptual categorization. Furthermore, we deliberately abstain from a unified quantitative benchmark, as the efficacy of guidance is deeply coupled with the rapidly evolving capabilities of base models; techniques essential for weaker models may become redundant as reasoning scales, rendering static comparisons scientifically fragile. Thus, this work serves as a qualitative synthesis of representative control paradigms rather than an exhaustive empirical leaderboard.

References

Ahmed Alajrami, Xingwei Tan, and Nikolaos Aletras. 2025. Fine-tuning on noisy instructions: Effects on generalization and performance. *Preprint*, arXiv:2510.03528.

Omer Antverg, Eyal Ben-David, and Yonatan Belinkov. 2022. IDANI: Inference-time domain adaptation via neuron-level interventions. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 21–29, Hybrid. Association for Computational Linguistics.

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2025. Safeinfer: Context adaptive decoding time safety alignment for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27188–27196.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: solving elaborate problems with large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zouying Cao, Yifei Yang, and Hai Zhao. 2025. Scans: mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. RQ-RAG: Learning to refine queries for retrieval augmented generation. In *First Conference on Language Modeling*.

Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, and Derek F. Wong. 2025a. SGIC: A self-guided iterative calibration framework for RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28357–28370, Vienna, Austria. Association for Computational Linguistics.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2025b. PAD: Personalized alignment at decoding-time. In *The Thirteenth International Conference on Learning Representations*.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, YiQiu Guo, Yanfeng Wang, and Yu Wang. 2025c. Towards omni-RAG: Comprehensive retrieval-augmented generation for large language models in medical applications. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15285–15309, Vienna, Austria. Association for Computational Linguistics.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024a. Black-box prompt optimization: Aligning large language models without model training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3201–3219, Bangkok, Thailand. Association for Computational Linguistics.

Kewei Cheng, Nesreen K. Ahmed, Theodore L. Willke, and Yizhou Sun. 2024b. Structure guided prompt: Instructing large language model in multi-step reasoning by exploring graph structure of the text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9407–

9430, Miami, Florida, USA. Association for Computational Linguistics.

Rong Cheng, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. 2025. DualRAG: A dual-process approach to integrate reasoning and retrieval for multi-hop question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31877–31899, Vienna, Austria. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*. Key foundational work on SAEs.

Hannah Cyberey and David Evans. 2025. Steering the censorship: Uncovering representation vectors for LLM "thought" control. In *Second Conference on Language Modeling*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Van Dai Do, Quan Hung Tran, Svetha Venkatesh, and Hung Le. 2025. Dynamic steering with episodic memory for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13731–13749, Vienna, Austria. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Jinyuan Fang, Zaiqiao Meng, and Craig MacDonald. 2025a. KiRAG: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18969–18985, Vienna, Austria. Association for Computational Linguistics.

Yuanheng Fang, Guoqing Chao, Wenqiang Lei, Shaobo Li, and Dianhui Chu. 2025b. Cdw-cot: clustered distance-weighted chain-of-thoughts reasoning. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Farima Fatahi Bayat, Xin Liu, H. Jagadish, and Lu Wang. 2024. Enhanced language model truthfulness with learnable intervention and uncertainty expression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12388–12400, Bangkok, Thailand. Association for Computational Linguistics.

Yu Fei, Yasaman Razeghi, and Sameer Singh. 2025. Nudging: Inference-time alignment of llms via guided decoding. *Preprint*, arXiv:2410.09300.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2024. Promptbreeder: self-referential self-improvement via prompt evolution. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Tingchen Fu, Yupeng Hou, Julian McAuley, and Rui Yan. 2025. Unlocking decoding-time controllability: Gradient-free multi-objective alignment with contrastive prompts. 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 366–384, Albuquerque, New Mexico. Association for Computational Linguistics.

Jun Gao, Qi Lv, Zili Wang, Tianxiang Wu, Ziqiang Cao, and Wenjie Li. 2025. UniICL: An efficient ICL framework unifying compression, selection, and generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 500–510, Vienna, Austria. Association for Computational Linguistics.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, and 1 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638. Published online 17 September 2025.

Dongyoon Hahm, Woogyeol Jin, June Suk Choi, Sungsoo Ahn, and Kimin Lee. 2025. Enhancing LLM agent safety via causal influence prompting. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15143–15168, Vienna, Austria. Association for Computational Linguistics.

Jongwook Han, Jongwon Lim, Injin Kong, and Yohan Jo. 2025. Dual mechanisms of value expression: Intrinsic vs. prompted values in llms. *Preprint*, arXiv:2509.24319.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. 2025. Mapping post-training forgetting in language models at scale. *Preprint*, arXiv:2510.17776.

Yinhan He, Wendy Zheng, Yaochen Zhu, Zaiyi Zheng, Lin Su, Sriram Vasudevan, Qi Guo, Liangjie Hong, and Jundong Li. 2025a. Semcot: Accelerating chain-of-thought reasoning through semantically-aligned implicit tokens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zeqing He, Zhibo Wang, Huiyu Xu, Hejun Lin, Wenhui Zhang, and Zhixuan Chu. 2025b. Interpretable LLM guardrails via sparse representation steering. *arXiv preprint arXiv:2503.16851*.

Amr Hegazy, Mostafa Elhoushi, and Amr Alanwar. 2025. Guiding giants: Lightweight controllers for weighted activation steering in llms. *Preprint*, arXiv:2505.20309.

Jakub Hoscilowicz, Adam Wiacek, Jan Chojnacki, Adam Cieslak, Leszek Michon, Vitalii Urbanevych, and Artur Janicki. 2024. Non-linear inference time intervention: Improving llm truthfulness. *Preprint*, arXiv:2403.18680.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Senkang Hu, Xudong Han, Jinqi Jiang, Yihang Tao, Zihan Fang, Sam Tak Wu Kwong, and Yuguang Fang. 2025a. Distribution-aligned decoding for efficient LLM task adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiaomeng Hu, Fei Huang, Chenhan Yuan, Junyang Lin, and Tsung-Yi Ho. 2025b. CARE: Decoding-time safety alignment via rollback and introspection intervention. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. 2025a. How to enable effective cooperation between humans and NLP models: A survey of principles, formalizations, and beyond. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 466–488, Vienna, Austria. Association for Computational Linguistics.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. 2025b. DeAL: Decoding-time alignment for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26280–26300, Vienna, Austria. Association for Computational Linguistics.

Ko-Wei Huang, Yi-Fu Fu, Ching-Yu Tsai, Yu-Chieh Tu, Tzu-ling Cheng, Cheng-Yu Lin, Yi-Ting Yang, Heng-Yi Liu, Keng-Te Liao, Da-Cheng Juan, and Shou-De Lin. 2025c. Neuron-level differentiation of memorization and generalization in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16077–16091, Suzhou, China. Association for Computational Linguistics.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Haozhe Ji, Pei Ke, Hongning Wang, and Minlie Huang. 2024. Language model decoding as direct metrics optimization. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Chonghe Jiang, Bao Nguyen, Anthony Man-Cho So, and Viet Anh Nguyen. 2025a. Probe-free low-rank activation intervention. 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 2812–2824, Albuquerque, New Mexico. Association for Computational Linguistics.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023a. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.

Xinke Jiang, Yue Fang, Rihong Qiu, Haoyu Zhang, Yongxin Xu, Hao Chen, Wentao Zhang, Ruizhe Zhang, Yuchen Fang, Xinyu Ma, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025b. TC–RAG: Turing–complete RAG's case study on medical LLM systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11400–11426, Vienna, Austria. Association for Computational Linguistics.

Xinke Jiang, Ruizhe Zhang, Yongxin Xu, Rihong Qiu, Yue Fang, Zhiyuan Wang, Jinyi Tang, Hongxin Ding, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025c. HyKGE: A hypothesis knowledge graph enhanced RAG framework for accurate and reliable medical LLMs responses. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11856, Vienna, Austria. Association for Computational Linguistics.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *Preprint*, arXiv:2312.03813.

Sathwik Karnik and Somil Bansal. 2025. Preemptive detection and steering of llm misalignment via latent reachability. *Preprint*, arXiv:2509.21528.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Minbeom Kim, Kang-il Lee, Seongho Joo, Hwaran Lee, Thibaut Thonet, and Kyomin Jung. 2025. Drift: Decoding-time personalized alignments with implicit user preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6107–6126, Suzhou, China. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. 2024. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, St. Julian's, Malta. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sachin Kumar, Yulia Tsvetkov, and 1 others. 2025. *Foundations of Large Language Models*. arXiv preprint arXiv:2501.09223.

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2771–2791, Vienna, Austria. Association for Computational Linguistics.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025a. Programming refusal with conditional activation steering. In *The Thirteenth International Conference on Learning Representations*.

Hakyung Lee, Subeen Park, Joowang Kim, Sungjun Lim, and Kyungwoo Song. 2025b. Uncertainty-aware contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26376–26391, Vienna, Austria. Association for Computational Linguistics.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022. Generative multi-hop retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Bolian Li, Yifan Wang, Anamika Lochab, Ananth Grama, and Ruqi Zhang. 2025a. Cascade reward sampling for efficient decoding-time alignment. *Preprint*, arXiv:2406.16306.

Dong Li, Yichen Niu, Ying Ai, Xiang Zou, Biqing Qi, and Jianxing Liu. 2025b. T-grag: A dynamic graphrag framework for resolving temporal conflicts and redundancy in knowledge retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11880–11889, New York, NY, USA. Association for Computing Machinery.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Pingzhi Li, Zhen Tan, Mohan Zhang, Huaizhi Qu, Huan Liu, and Tianlong Chen. 2025c. DOGe: Defensive output generation for LLM protection against knowledge distillation. *arXiv preprint arXiv:2505.19504*.

Qiwei Li, Teng Xiao, Zuchao Li, Ping Wang, Mengjia Shen, and Hai Zhao. 2025d. Dialogue-RAG: Enhancing retrieval for LLMs via node-linking utterance rewriting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24423–24438, Vienna, Austria. Association for Computational Linguistics.

Rui Li, Liyang He, Qi Liu, Zheng Zhang, Heng Yu, Yuyang Ye, Linbo Zhu, and Yu Su. 2025e. UniRAG: Unified query understanding method for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14163–14178, Vienna, Austria. Association for Computational Linguistics.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024b. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Baohao Liao, Yuhui Xu, Hanze Dong, Junnan Li, Christof Monz, Silvio Savarese, Doyen Sahoo, and Caiming Xiong. 2025. Reward-guided speculative decoding for efficient LLM reasoning. In *Forty-second International Conference on Machine Learning*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Jiaen Lin, Jingyu Liu, and Yingbo Liu. 2025a. Optimizing multi-hop document retrieval through intermediate representations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15798–15809, Vienna, Austria. Association for Computational Linguistics.

Yukang Lin, Bingchen Zhong, Shuoran Jiang, Joanna Siebert, and Qingcai Chen. 2025b. Reasoning graph enhanced exemplars retrieval for in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9737–9759, Abu Dhabi, UAE. Association for Computational Linguistics.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025a. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1897–1913, Vienna, Austria. Association for Computational Linguistics.

Hui Liu, Wenya Wang, Hao Sun, Chris Xing Tian, Chenqi Kong, Xin Dong, and Haoliang Li. 2025b. Unraveling the mechanics of learning-based demonstration selection for in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2623–2641, Vienna, Austria. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025c. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 2781–2790, New York, NY, USA. Association for Computing Machinery.

Sheng Liu, Tianlang Chen, Pan Lu, Haotian Ye, Yizheng Chen, Lei Xing, and James Zou. 2025d. Fractional reasoning via latent steering vectors improves inference time compute. *Preprint*, arXiv:2506.15882.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024b. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Tianlin Liu, Shangmin Guo, Leonardo Bianco, Daniele Calandriello, Quentin Berthet, Felipe Llinares, Jessica Hoffmann, Lucas Dixon, Michal Valko, and Mathieu Blondel. 2024c. Decoding-time realignment of language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2025. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Nicholas Matsumoto, Jay Moran, Hyunjun Choi, Miguel E Hernandez, Mythreye Venkatesan, Paul Wang, and Jason H Moore. 2024. Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models. *Bioinformatics*, 40(6):btae353.

Oscar Mañas, Pierluca D'Oro, Koustuv Sinha, Adriana Romero-Soriano, Michal Drozdzal, and Aishwarya Agrawal. 2025. Controlling multimodal llms via reward-guided decoding. *Preprint*, arXiv:2508.11616.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

Quang Duc Nguyen, Tung Nguyen, Duc Anh Nguyen, Linh Ngo Van, Sang Dinh, and Thien Huu Nguyen. 2025. GloCOM: A short text neural topic model via global clustering context. 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 1109–1124, Albuquerque, New Mexico. Association for Computational Linguistics.

Lin Ning, Harsh Lara, Meiqi Guo, and Abhinav Rastogi. 2025. MoDE: Effective multi-task parameter efficient fine-tuning with a mixture of dyadic experts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8233–8246, Albuquerque, New Mexico. Association for Computational Linguistics.

Narmeen Fatimah Oozeer, Dhruv Nathawani, Nirmalendu Prakash, Michael Lan, Abir HARRASSE, and Amir Abdullah. 2025. Activation space interventions can be transferred between large language models. In *Forty-second International Conference on Machine Learning*.

OpenAI and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Birong Pan, Yongqi Li, Weiyu Zhang, Wenpeng Lu, Mayi Xu, Shen Zhou, Yuanyuan Zhu, Ming Zhong, and Tieyun Qian. 2025. A survey on training-free alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4445–4461, Suzhou, China. Association for Computational Linguistics.

Yein Park, Jungwoo Park, and Jaewoo Kang. 2025. Asguard: Activation-scaling guard to mitigate targeted jailbreaking attack. *Preprint*, arXiv:2509.25843.

Mansi Phute, Alec Helbling, Matthew Daniel Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. LLM self defense: By self examination, LLMs know they are being tricked. In *The Second Tiny Papers Track at ICLR 2024*.

Joris Postmus and Steven Abreu. 2024. Steering large language models using conceptors: Improving addition-based activation engineering. In *MINT: Foundation Model Interventions*.

Zhenting Qi, Mingyuan MA, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2025a. Mutual reasoning makes smaller LLMs stronger problem-solver. In *The Thirteenth International Conference on Learning Representations*.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. 2025b. rstar: Mutual reasoning makes smaller LLMs stronger problem-solver. *arXiv preprint arXiv:2501.04519*. Microsoft Research.

Zackary Rackauckas. 2024. Rag-fusion: A new take on retrieval augmented generation. *International Journal on Natural Language Computing*, 13(1):37–47.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nate Rahn, Pierluca D'Oro, and Marc G Bellemare. 2024. Controlling large language model agents with entropic activation steering. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. A practical review of mechanistic interpretability for transformer-based language models. *Preprint*, arXiv:2407.02646.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Federico Raspanti, Tanir Ozcelebi, and Mike Holenderski. 2025. Grammar-constrained decoding makes large language models better logical parsers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 485–499, Vienna, Austria. Association for Computational Linguistics.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. *Preprint*, arXiv:2312.09300.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. 2025. SmoothLLM: Defending large language models against jailbreaking attacks. *Transactions on Machine Learning Research*.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. 2024. Controlling language and diffusion models by transporting activations. *Preprint*, arXiv:2410.23054.

Mansi Sakarvadia, Aswathy Ajith, Arham Mushtaq Khan, Nathaniel C Hudson, Caleb Geniesse, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W. Mahoney. 2025. Mitigating memorization in language models. In *The Thirteenth International Conference on Learning Representations*.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing*

and Interpreting Neural Networks for NLP, pages 577–603, Miami, Florida, US. Association for Computational Linguistics.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

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

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hannaneh Hajishirzi, Noah A. Smith, and Simon Shaolei Du. 2024a. Decoding-time language model alignment with multiple objectives. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024b. 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 8371–8384, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kashun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore. Association for Computational Linguistics.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Machine Learning (ICML)*.

Seongho Son, William Bankes, Sangwoong Yoon, Shyam Sundhar Ramesh, Xiaohang Tang, and Ilija Bogunovic. 2025. Robust multi-objective controlled decoding of large language models. *Preprint*, arXiv:2503.08796.

Feifan Song, Yuxuan Fan, Xin Zhang, Peiyi Wang, and Houfeng Wang. 2025a. Instantly learning preference alignment via in-context DPO. 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 161–178, Albuquerque, New Mexico. Association for Computational Linguistics.

Feifan Song, Bofei Gao, Yifan Song, Yi Liu, Weimin Xiong, Yuyang Song, Tianyu Liu, Guoyin Wang, and Houfeng Wang. 2025b. P-aligner: Enabling pre-alignment of language models via principled instruction synthesis. *Preprint*, arXiv:2508.04626.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. Interpretable steering of large language models with feature guided activation additions. *Preprint*, arXiv:2501.09929.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. 2024a. Fast best-of-n decoding via speculative rejection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. 2025a. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jiashuo Sun, Xianrui Zhong, Sizhe Zhou, and Jiawei Han. 2025b. DynamicRAG: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Simeng Sun, Yang Liu, Shuohang Wang, Dan Iter, Chenguang Zhu, and Mohit Iyyer. 2024b. PEARL: Prompting large language models to plan and execute actions over long documents. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 469–486, St. Julian's, Malta. Association for Computational Linguistics.

Yushi Sun, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2025c. KERAG: Knowledge-enhanced retrieval-augmented generation for advanced question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6194–6216, Suzhou, China. Association for Computational Linguistics.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2023. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4265–4293, Toronto, Canada. Association for Computational Linguistics.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *Preprint*, arXiv:2505.12189.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. 2024. Extending activation steering to broad skills and multiple behaviours. *Preprint*, arXiv:2403.05767.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025a. Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571, Vienna, Austria. Association for Computational Linguistics.

Han Wang, Gang Wang, and Huan Zhang. 2025b. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. *arXiv preprint arXiv:2411.16721*.

Liang Wang, Nan Yang, and Furu Wei. 2024a. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Mengru Wang, Ziwen Xu, Shengyu Mao, Shumin Deng, Zhaopeng Tu, Huajun Chen, and Ningyu Zhang. 2025c. Beyond prompt engineering: Robust behavior control in LLMs via steering target atoms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23381–23399, Vienna, Austria. Association for Computational Linguistics.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. 2025d. Adaptive activation steering: A tuning-free LLM truthfulness improvement method for diverse hallucinations categories. In *THE WEB CONFERENCE 2025*.

Weixuan Wang, JINGYUAN YANG, and Wei Peng. 2025e. Semantics-adaptive activation intervention for LLMs via dynamic steering vectors. In *The Thirteenth International Conference on Learning Representations*.

Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2024b. Promptagent: Strategic planning with language models enables expert-level prompt optimization. In *The Twelfth International Conference on Learning Representations*.

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

Xunguang Wang, Daoyuan Wu, Zhenlan Ji, Zongjie Li, Pingchuan Ma, Shuai Wang, Yingjiu Li, Yang Liu, Ning Liu, and Juergen Rahmel. 2025f. Selfdefend: Llms can defend themselves against jailbreaking in a practical manner. In *Proceedings of the 34th USENIX Conference on Security Symposium*, SEC '25, USA. USENIX Association.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024c. Factuality of

large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Wikipedia contributors. 2025a. Guidance, navigation, and control. https://en.wikipedia.org/wiki/Guidance%2C_navigation%2C_and_control. Accessed: 2025-12-01.

Wikipedia contributors. 2025b. Steering. https://en.wikipedia.org/wiki/Steering. Accessed: 2025-12-01.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. 2025a. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28443–28467, Vienna, Austria. Association for Computational Linguistics.

Lyucheng Wu, Mengru Wang, Ziwen Xu, Tri Cao, Nay Oo, Bryan Hooi, and Shumin Deng. 2025b. Automating steering for safe multimodal large language models. *Preprint*, arXiv:2507.13255.

Mingyan Wu, Zhenghao Liu, Yukun Yan, Xinze Li, Shi Yu, Zheni Zeng, Yu Gu, and Ge Yu. 2025c. RankCoT: Refining knowledge for retrieval-augmented generation through ranking chain-of-thoughts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12857–12874, Vienna, Austria. Association for Computational Linguistics.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. MolRAG: Unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531, Vienna, Austria. Association for Computational Linguistics.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.

Ran Xu, Hui Liu, Sreyashi Nag, Zhenwei Dai, Yaochen Xie, Xianfeng Tang, Chen Luo, Yang Li, Joyce C. Ho, Carl Yang, and Qi He. 2025. SimRAG: Self-improving retrieval-augmented generation for adapting large language models to specialized domains. 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 11534–11550, Albuquerque, New Mexico. Association for Computational Linguistics.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, and Yulan He. 2024. Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10423–10435, Miami, Florida, USA. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2025a. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. 2025b. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Jinghan Yang, Zhenbo Xu, Dehua Ma, Liu Liu, Fei Liu, Gong Huang, and Zhaofeng He. 2025c. Reciperag: Advancing recipe generation with reinforced retrieval augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 5060–5069, New York, NY, USA. Association for Computing Machinery.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Zairun Yang, Yilin Wang, Zhengyan Shi, Yuan Yao, Lei Liang, Keyan Ding, Emine Yilmaz, Huajun Chen, and Qiang Zhang. 2025d. EventRAG: Enhancing LLM generation with event knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 16967–16979, Vienna, Austria. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Hao Ye, Mengshi Qi, Zhaohong Liu, Liang Liu, and Huadong Ma. 2025. Safedriverag: Towards safe autonomous driving with knowledge graph-based retrieval-augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11170–11178, New York, NY, USA. Association for Computing Machinery.

Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. 2024. Sled: self logits evolution decoding for improving factuality in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025a. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882, Vienna, Austria. Association for Computational Linguistics.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. 2025b. A survey on test-time scaling in large language models: What, how, where, and how well? *Preprint*, arXiv:2503.24235.

Xiaoyun Zhang, Zhengyue Zhao, Wenxuan Shi, Kaidi Xu, Di Huang, and Xing Hu. 2025c. Safety alignment of large language models via contrasting safe and harmful distributions. *Preprint*, arXiv:2406.16743.

Yuqi Zhang, Liang Ding, Lefei Zhang, and Dacheng Tao. 2025d. Intention analysis makes LLMs a good jailbreak defender. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2947–2968, Abu Dhabi, UAE. Association for Computational Linguistics.

Zheng Zhang, Shaocheng Lan, Lei Song, Jiang Bian, Yexin Li, and Kan Ren. 2025e. Learning to select in-context demonstration preferred by large language model. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11345–11360, Vienna, Austria. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhuoxuan Zhang, Jinhao Duan, Edward Kim, and Kaidi Xu. 2025f. Sparse neurons carry strong signals of question ambiguity in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16092–16110, Suzhou, China. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025a. A survey of large language models. *Preprint*, arXiv:2303.18223.

Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. 2024a. SEER: Self-aligned evidence extraction for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3041, Miami, Florida, USA. Association for Computational Linguistics.

Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. 2024b. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6144–6166, Torino, Italia. ELRA and ICCL.

Zekai Zhao, Qi Liu, Kun Zhou, Zihan Liu, Yifei Shao, Zhiting Hu, and Biwei Huang. 2025b. Activation control for efficiently eliciting long chain-of-thought ability of language models. *Preprint*, arXiv:2505.17697.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2024. ROSE doesn't do that: Boosting the safety of instruction-tuned large language models with reverse prompt contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13721–13736, Bangkok, Thailand. Association for Computational Linguistics.

Tianqi Zhong, Quan Wang, Jingxuan Han, Yongdong Zhang, and Zhendong Mao. 2023. Air-decoding: Attribute distribution reconstruction for decoding-time controllable text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8233–8248, Singapore. Association for Computational Linguistics.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2024a. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the*

41st International Conference on Machine Learning, ICML'24. JMLR.org.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023b. Large language models are human-level prompt engineers. *Preprint*, arXiv:2211.01910.

Yujun Zhou, Yufei Han, Haomin Zhuang, Kehan Guo, Zhenwen Liang, Hongyan Bao, and Xiangliang Zhang. 2024b. Defending jailbreak prompts via in-context adversarial game. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20084–20105, Miami, Florida, USA. Association for Computational Linguistics.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1615–1624.

Jin Ziqi and Wei Lu. 2023. Tab-CoT: Zero-shot tabular chain of thought. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10259–10277, Toronto, Canada. Association for Computational Linguistics.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Scope and Rationale of this Survey

Rationale: Unifying a Fragmented Landscape. As illustrated in Table 1, the field of LLM adaptation is undergoing a fundamental shift from parameter-heavy training to flexible test-time interventions. However, the current literature treats these interventions as distinct, unrelated fields. Researchers working on *Prompt Engineering* (Input Space) often operate independently from those studying *Mechanistic Interpretability* (Latent Space) or *Constrained Decoding* (Decoding Space). Existing surveys typically focus either on isolated subfields (Schulhoff et al., 2025; Rai et al., 2025) or on specific objectives like safety (Pan et al., 2025) or efficiency (Zhang et al., 2025b), obscuring the underlying control mechanisms. This survey is essential to bridge these silos. By proposing the first mechanism-centric taxonomy, we reveal how the same mathematical principles can be applied across different stages of generation, providing researchers with a "full-stack" view of model controllability that existing goal-oriented reviews lack.

Scope: Defining Training-Free Guidance. To ensure clarity, we strictly define Training-Free Test-Time Guidance based on two criteria: (1) Frozen Parameters, meaning the parameters of the backbone LLM are strictly frozen and cannot be trained; and (2) Inference Intervention, where the method actively alters the model's generation probability or trajectory to satisfy specific constraints. Under this definition, our review covers the entire inference lifecycle: Input Space, Latent Space, Decoding Space, and Output Space. We explicitly exclude system-level optimizations focused solely on throughput and permanent weight-editing techniques, as they aim for computational efficiency rather than control.

B Terminology: Why We Use the Term "Guidance"

In this survey, we use *guidance* as an umbrella term for LLM-training-free, test-time mechanisms that inject additional signals into the inference process (in the input space, latent space, or decoding process) to direct a frozen model toward a behavioural objective, without updating its parameters. This notion is close to the role of a guidance module in control systems, which specifies the desired trajectory of a system given its current state and mission goal, distinct from both navigation and low-level control (Wikipedia contributors, 2025a).

We deliberately distinguish *guidance* from several related terms. *Steering* typically denotes changing the direction of motion (Wikipedia contributors, 2025b) and, in the LLM literature, is often used for activation-level representation engineering—adding or modifying directions in hidden space to control behaviour (Zou et al., 2025). *Interven-*

Table 1: Comparison with existing surveys related to LLM adaptation. ✔ = Primary focus with detailed taxonomy; ● = Partial coverage or sub-topic discussion; ✗ = Little to no coverage. **Our survey** is the first to provide a unified framework covering the entire test-time inference lifecycle.

Survey	Core Focus	Intervention Space (Lifecycle)				Unified View
		Input	Latent	Decoding	Output	
General & Training-Based Surveys						
Zhao et al. (Zhao et al., 2025a)	General LLM Capabilities	●	✗	✗	●	✗
Tie et al. (Lai et al., 2025)	Post-Training Scaling	✗	✗	✗	✗	✗
Subfield-Specific Surveys						
Schulhoff et al. (Schulhoff et al., 2025)	Prompt Engineering Techniques	✔	✗	✗	✗	✗
Fan et al. (Fan et al., 2024)	Retrieval-Augmented Generation	✔	✗	✗	✗	✗
Rai et al. (Rai et al., 2025)	Mechanistic Interpretability	✗	✔	✗	✗	✗
Goal-Oriented Test-Time Surveys						
Zhang et al. (Zhang et al., 2025b)	Test-Time Scaling	●	✗	●	✔	●
Pan et al. (Pan et al., 2025)	Training-Free Alignment	✔	●	✔	●	✗
Huang et al. (Huang et al., 2025a)	Human-AI Cooperation	✔	✗	✗	●	✗
Ours	**Test-Time Guidance Mechanism**	✔	✔	✔	✔	✔

tion usually emphasises the concrete act of editing internal activations at specific layers or tokens during inference (Li et al., 2023a). *Mitigation* describes a safety-oriented objective (e.g., reducing hallucinations or toxicity), rather than a particular mechanism.

In this paper, we therefore use *guidance* as the method-agnostic, trajectory-oriented term for test-time control of LLMs; we treat *steering* and *interventions* as specific implementation families within this space, and *mitigation* as one important class of guidance objectives (safety and robustness).

C A Unified Guidance Theoretical Framework from a Conditional Distribution Shaping Perspective

To synthesize the diverse methods categorized in our taxonomy, we propose a unified framework based on Conditional Distribution Shaping (Ji et al., 2024; Hu et al., 2025a; Li et al., 2025c). Fundamentally, a pre-trained LLM defines a base autoregressive distribution $P_\theta(y|x)$. The objective of any training-free guidance method is to construct a transformed distribution $\tilde{P}(y|x;g)$ that satisfies specific constraints imposed by the guidance signal g.

We formalize this transformation as the application of a shaping operator $\mathcal{T}$ at a specific interface $\mathcal{I}$ of the model's computational graph (as illustrated in Figure 6):

$$\tilde{P}(y|x;g) = \mathcal{T}_{\mathcal{I}}\left(P_\theta(\cdot|x), g\right) \tag{8}$$

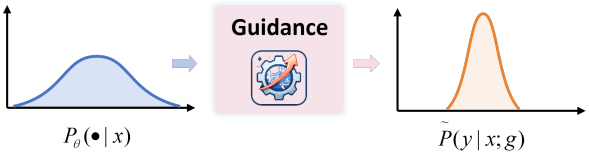

Figure 6: Schematic of the Unified Guidance Theoretical Framework from a Conditional Distribution Shaping Perspective.

The four intervention spaces in our taxonomy correspond to applying $\mathcal{T}$ at four distinct depth levels of the generative process.

1. Input Space ($\mathcal{T}_{\text{input}}$): In input-level guidance, the shaping operator intervenes on the conditioning variable itself. Rather than modifying the model's mapping function, it seeks a perturbation δ in the discrete token space (or embedding space) such that the base model naturally projects the input to the desired output region.

$$\tilde{P}(y|x;g) = P_\theta(y \mid \phi(x,g)) \tag{9}$$

where $\phi(x,g)$ represents the transformation function (e.g., prepending a system prompt or retrieving context). Theoretically, this leverages the pre-training manifold continuity: it relies on the assumption that there exists a neighborhood $\phi(x,g)$

in the context space that maps to the target distribution $\tilde{P}$ solely through the frozen parameters θ.

2. Latent Space ($\mathcal{T}_{\text{latent}}$): Latent interventions operate by warping the intermediate feature mapping. Let $F_\ell(h_{\ell-1})$ be the transformation at layer ℓ. Guidance introduces an affine transformation (steering vector v) to the hidden state h_ℓ. The shaped distribution is the result of forward-propagating this perturbed state:

$$\tilde{P}(y|x;g) = \text{Head}\left(\dots F_L(h_\ell + \lambda v_g)\dots\right) \quad (10)$$

Unlike input guidance which is discrete, this operator performs a geometric translation on the semantic manifold. By shifting the representation h_ℓ along the direction v_g, the operator $\mathcal{T}_{\text{latent}}$ biases the subsequent layers to process the context as if it possessed the attribute g, effectively reshaping the downstream probability surface before it is projected to the vocabulary.

3. Decoding Space ($\mathcal{T}_{\text{decoding}}$): Decoding guidance acts directly on the output probability mass function at each step t. The operator $\mathcal{T}$ transforms the logits z_t into $\tilde{z}_t$ via an additive or masking term $\psi(g)$, followed by re-normalization:

$$\tilde{P}(y_t|y_{<t},x;g) = \text{Softmax}\left(z_t + \lambda \cdot \psi(y_t,g)\right) \quad (11)$$

Here, ψ can be a constraint mask or a contrastive bias. This represents a local re-weighting strategy. The shaping operator explicitly "carves" the distribution, suppressing the probability mass of tokens that violate g and re-allocating it to compliant tokens, ensuring strict or soft adherence to constraints at the token level.

4. Output Space ($\mathcal{T}_{\text{output}}$): Output guidance operates on the distribution over complete sequences. Theoretically, this can be viewed as applying a selection filter over the support set $\mathcal{Y}$. The shaped distribution becomes:

$$\tilde{P}(y|x;g) = \frac{P_\theta(y|x)\cdot I[R(y,g)>\tau]}{Z} \quad (12)$$

where I is an indicator function (or soft score) from a verifier R, and Z is the normalization constant. This operator performs global pruning. It shapes the distribution by truncating the tail and concentrating the probability mass solely on the candidate y^* that maximizes the global verification score.

D Literature Search Methodology

To ensure a comprehensive coverage of the landscape, we conducted a systematic search across major academic databases, including ACL Anthology, Google Scholar, and Semantic Scholar Complementing these traditional repositories, we actively monitored real-time research trends using AI-driven discovery platforms such as PaSa[1], Papers.cool[2], and Hugging Face Papers[3]. We also employed a "snowballing" strategy, leveraging bibliographies from existing related surveys (as discussed in Table 1) to identify relevant training-free mechanisms. We strictly filtered papers based on the criteria of keeping the backbone model frozen and involving active inference-time intervention. The complete consolidated list of surveyed literature, categorized by intervention space, is visualized in Figure 7.

[1] https://pasa-agent.ai/
[2] https://papers.cool/
[3] https://huggingface.co/papers/trending

I. Input-Space Guidance

Contextual Augmentation: IC-RAG (Ram et al., 2023), REPLUG (Shi et al., 2024b), CoK (Li et al., 2024b), RQ-RAG (Chan et al., 2024), GMR (Lee et al., 2022), RAG-Fusion (Rackauckas, 2024), KRAGEN (Matsumoto et al., 2024), SimRAG (Xu et al., 2025), SEER (Zhao et al., 2024a), L-RAG (Lin et al., 2025a), HopRAG (Liu et al., 2025a), KERAG (Sun et al., 2025c), TC-RAG (Jiang et al., 2025b), HyKGE (Jiang et al., 2025c), UniRAG (Li et al., 2025e), OmniRAGMed (Chen et al., 2025c), MolRAG (Xian et al., 2025), EventRAG (Yang et al., 2025d), KiRAG (Fang et al., 2025a), FaithfulRAG (Zhang et al., 2025a), Dialogue-RAG (Li et al., 2025d), SGIC (Chen et al., 2025a), MedGraphRAG (Wu et al., 2025a), Astute RAG (Wang et al., 2025a), DualRAG (Cheng et al., 2025), T-GRAG (Li et al., 2025b), SafeDriveRAG (Ye et al., 2025), RecipeRAG (Yang et al., 2025c), HM-RAG (Liu et al., 2025c), DynamicRAG (Sun et al., 2025b), What-makes-ICL (Liu et al., 2022), Lost-in-Middle (Liu et al., 2024a), L2R-ICL (Wang et al., 2024a), CED-Select (Iter et al., 2023), RGER (Lin et al., 2025b), GENICL (Zhang et al., 2025e), MLSM/TTF (Liu et al., 2025b), UniICL (Gao et al., 2025), FLARE (Jiang et al., 2023b), CoVe (Wang et al., 2024c), ToG2 (Ma et al., 2025).

Prompt Design and Optimization: Chain-of-Thought (Wei et al., 2022), Zero-shot CoT (Kojima et al., 2022), Self-Consistency (Wang et al., 2023), Tab-CoT (Ziqi and Lu, 2023), Auto-CoT (Zhang et al., 2023), Logical CoT (Zhao et al., 2024b), CDW-CoT (Fang et al., 2025b), RankCoT (Wu et al., 2025c), PCoT (Modzelewski et al., 2025), SemCoT (He et al., 2025a), APE (Zhou et al., 2023b), OPRO (Yang et al., 2024), BPO (Cheng et al., 2024a), RLPrompt (Deng et al., 2022), PromptAgent (Wang et al., 2024b), PEARL-plan (Sun et al., 2024b), PromptBreeder (Fernando et al., 2024), Structure-Guided (Cheng et al., 2024b), AutoPrompt (Shin et al., 2020), Automate-CoT (Shum et al., 2023).

Alignment and Safety Prompting: Self-Reminders (Xie et al., 2023), URIAL (Lin et al., 2024), LLaMA-2 cfg. (Touvron et al., 2023), ICAG (Zhou et al., 2024b), SmoothLLM (Robey et al., 2025), ROSE (Zhong et al., 2024), Instructive Decoding (Alajrami et al., 2025), Intention Analysis (Zhang et al., 2025d), CIP (Hahm et al., 2025), LLM-Self-Defense (Phute et al., 2024), SelfDefenD (Wang et al., 2025f), P-Aligner (Song et al., 2025b), ICDPO (Song et al., 2025a).

II. Latent-Space Guidance

Global Directional Arithmetic: Activation Engineering (Turner et al., 2024), PPLM (Dathathri et al., 2020), RepE (Zou et al., 2023), ITI (Li et al., 2023a), In-Context Vectors (Liu et al., 2024b), Function Vectors (Todd et al., 2024), Latent steering vectors (Subramani et al., 2022), IDANI (Antverg et al., 2022), Scaling Laws (Stoehr et al., 2024), Layer Sensitivity (Stolfo et al., 2025), Fractional Reasoning (Liu et al., 2025d).

Decomposition and Fine-Grained Control: SAE-based Steering (Huben et al., 2024; Yan et al., 2024), Interpretable Steering (Soo et al., 2025), Sparse Neurons (Zhang et al., 2025f), Neuron Dampening (Huang et al., 2025c), Dual Mechanisms (Han et al., 2025), Steering Target Atoms (Wang et al., 2025c), Concept Clamping (Valentino et al., 2025; Sakarvadia et al., 2025).

Geometric Constraints and Projection: AlphaSteer (Sheng et al., 2025), ASGuard (Park et al., 2025), Projection Steering (Postmus and Abreu, 2024), FLORAIN (Jiang et al., 2025a), Low-Rank Interventions (Oozeer et al., 2025), Diffusion Control (Rodriguez et al., 2024).

Context-Adaptive Dynamics: NL-ITI (Hoscilowicz et al., 2024), EAST (Rahn et al., 2024), SADI (Wang et al., 2025e), Adaptive Steering (Wang et al., 2025d), SCANS (Cao et al., 2025), Preemptive Detection (Karnik and Bansal, 2025), AutoSteer (Multimodal) (Wu et al., 2025b), Layer-Navigator (Sun et al., 2025a), DSAS (Do et al., 2025).

III. Decoding-Space Guidance

Score-based Reweighting: GeDi (Krause et al., 2021), DExperts (Liu et al., 2021), Contrastive Decoding (Li et al., 2023b), DoLa (Chuang et al., 2024), UACD (Lee et al., 2025b), SLED (Zhang et al., 2024), ROSE (Zhong et al., 2024), Instructive Decoding (Alajrami et al., 2025), ACD (Zhang et al., 2025c), MOD (Shi et al., 2024a), DeAL (Huang et al., 2025b), PAD (Chen et al., 2025b), Drift (Kim et al., 2025), SAFEINFER (Banerjee et al., 2025), CDT (Yang et al., 2025b), RMOD (Son et al., 2025), IBD (Zhu et al., 2024), Nudging (Fei et al., 2025), DeRa (Liu et al., 2024c), MCA (Fu et al., 2025), Air-Decoding (Zhong et al., 2023), FUDGE (Yang and Klein, 2021).

Constraint-based Filtering: NeuroLogic Decoding (Lu et al., 2021), Grammar-Constrained (Raspanti et al., 2025), CARE (Hu et al., 2025b), CARDS (Li et al., 2025a), RSD (Liao et al., 2025), RGD (Mañas et al., 2025), PRGD (Nguyen et al., 2025).

IV. Output-Space Guidance

Search and Planning: ToT (Yao et al., 2023a), GoT (Besta et al., 2024), RAP (Hao et al., 2023), Multi-Agent Debate/LATS (Zhou et al., 2024a).

Iterative Refinement: Reflexion (Shinn et al., 2023), Self-Refine (Madaan et al., 2023), CRITIC (Gou et al., 2024).

Candidate Selection and Reranking: Training Verifiers (Cobbe et al., 2021), SelfCheckGPT (Manakul et al., 2023), MBR Decoding (Suzgun et al., 2023), Universal Self-Consistency (Chen et al., 2024), Speculative Rejection (Sun et al., 2024a), Self-Evaluation (Ren et al., 2023).

Multi-Model Collaboration: COOPERATE (Feng et al., 2024), LLM-Blender (Jiang et al., 2023a), Selection-Inference (Creswell et al., 2023), Mutual Reasoning (Qi et al., 2025a).

Figure 7: A comprehensive summary of the surveyed literature, organized by the taxonomy detailed in Section 3.