A Survey on Training-free Alignment of Large Language Models

Birong Pan[1], **Yongqi Li**[1], **Weiyu Zhang**[2,*], **Wenpeng Lu**[2,*],
Mayi Xu[1], **Shen Zhou**[1], **Yuanyuan Zhu**[1], **Ming Zhong**[1], **Tieyun Qian**[1,3,*]
[1]School of Computer Science, Wuhan University, Wuhan, China
[2]Faculty of Computer Science and Technology, Qilu University of Technology, Shandong, China
[3]Zhongguancun Academy, Beijing, China
{panbirong,qty}@whu.edu.cn

Abstract

The alignment of large language models (LLMs) aims to ensure their outputs adhere to human values, ethical standards, and legal norms. Traditional alignment methods often rely on resource-intensive fine-tuning (FT), which may suffer from knowledge degradation and face challenges in scenarios where the model accessibility or computational resources are constrained. In contrast, training-free (TF) alignment techniques–leveraging in-context learning, decoding-time adjustments, and post-generation corrections–offer a promising alternative by enabling alignment without heavily retraining LLMs, making them adaptable to both open-source and closed-source environments. This paper presents the first systematic review of TF alignment methods, categorizing them by stages of **pre-decoding**, **in-decoding**, and **post-decoding**. For each stage, we provide a detailed examination from the viewpoint of LLMs and multimodal LLMs (MLLMs), highlighting their mechanisms and limitations. Furthermore, we identify key challenges and future directions, paving the way for more inclusive and effective TF alignment techniques. By synthesizing and organizing the rapidly growing body of research, this survey offers a guidance for practitioners and advances the development of safer and more reliable LLMs.

1 Introduction

The advent of Large Language Models (LLMs) (Hurst et al., 2024; Achiam et al., 2023; Touvron et al., 2023; Chiang et al., 2023) and Multimodal Large Language Models (MLLMs) (Bai et al., 2023; Chen et al., 2025; Ye et al., 2024) has marked a paradigm shift in human-machine interaction, enabling tasks ranging from complex reasoning to cross-modal understanding.

However, as these models permeate critical domains such as healthcare, education, and public discourse, their widespread adoption has ignited a dual-edged societal response: urgent concerns over risks and growing demands for enhanced capabilities. On one hand, LLMs bring about various risks like generating harmful content (Soice et al., 2023), perpetuating biases and compromising privacy (Salecha et al., 2024; Staab et al., 2024; Abdulhai et al., 2024). On the other hand, users and industries increasingly demand models that adapt to dynamic knowledge (Zhang et al., 2023; Si et al., 2023; Beukman et al., 2023), excel in specialized tasks (Li et al., 2024a; Chen et al., 2024a), and deliver personalized services (Richardson et al., 2023; Zhuang et al., 2024). These dual imperatives, namely mitigating risks and fulfilling functional demands, underscore the necessity of aligning LLMs with human values, ethical standards, and practical requirements. Alignment is no longer a technique alone but a prerequisite for responsible and effective deployment.

As a critical problem in artificial intelligence and natural language processing, LLM alignment has attracted considerable research attention (Shen et al., 2023; Wang et al., 2024d; Gabriel, 2020; Ouyang et al., 2022). Traditional alignment methods primarily rely on fine-tuning (FT), where model parameters are adjusted using curated datasets. While effective in specific scenarios, these approaches face one or more of the following three critical limitations: (1) *Severe Knowledge Degradation*: FT may overwrite pretrained knowledge, leading to catastrophic forgetting of general capabilities. (2) *Resource Intensity*: Collecting high-quality alignment data and retraining LLMs requires prohibitive computational and human resources. (3) *Access Constraint*: Proprietary models (e.g., GPT-4, Gemini) restrict parameter access, rendering FT infeasible.

To address these challenges, training-free (TF) alignment has emerged as a versatile paradigm. Instead of modifying model parameters, TF meth-

* Corresponding authors.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4445–4461

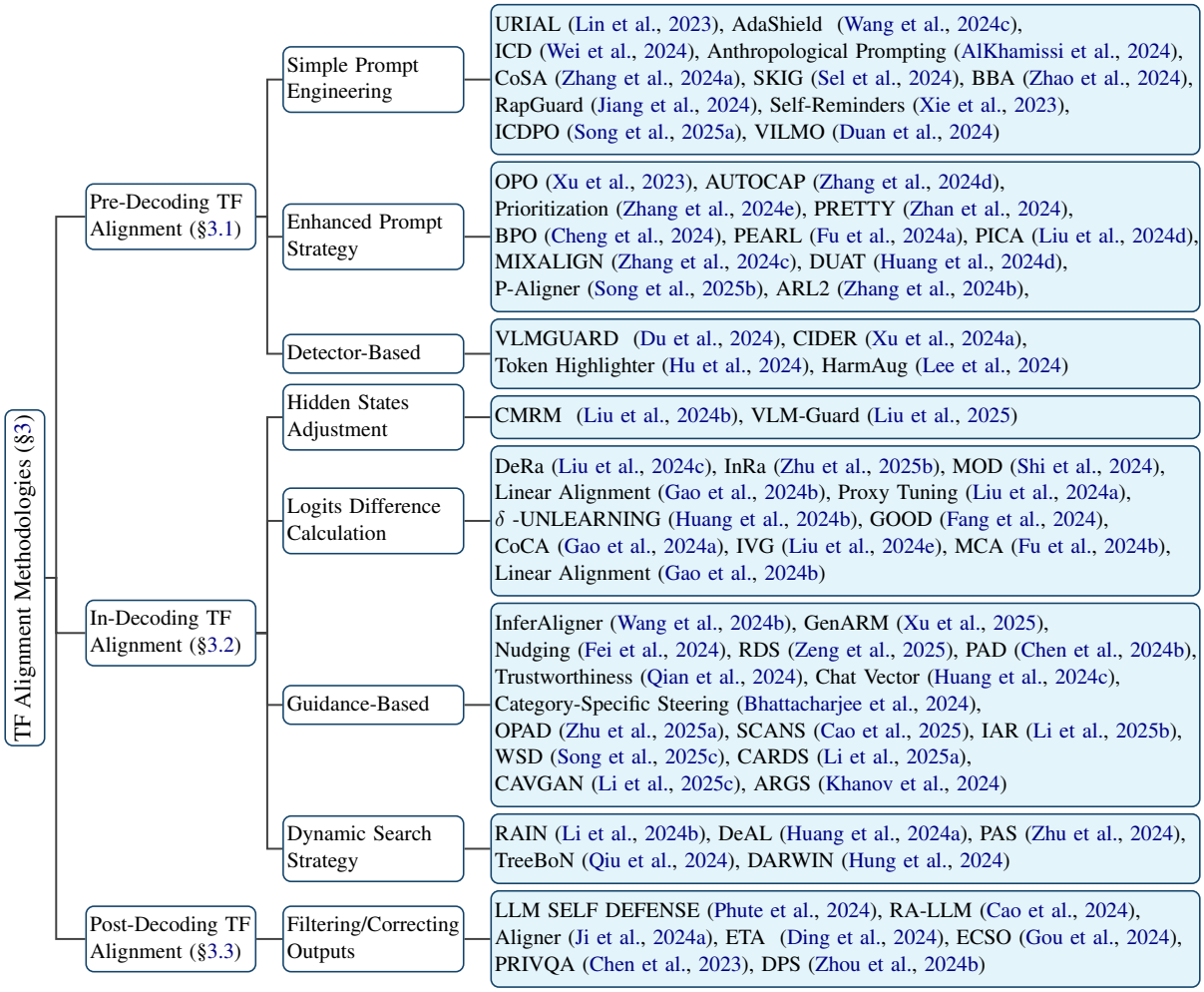

Figure 1: Taxonomy of training-free (TF) alignment methodologies for LLMs, categorized into pre-decoding, in-decoding, and post-decoding strategies.

ods intervene at different stages of the generation pipeline: (1) *Pre-Decoding TF Alignment*: guiding models by modifying inputs or prompts. (2) *In-Decoding TF Alignment*: adjusting token selection during generation. (3) *Post-Decoding TF Alignment*: filtering or refining outputs to meet safety and utility criteria. The overview of TF alignment methods at different stages is shown in Figure 1, which helps to gain a comprehensive understanding of the process.

While TF alignment demonstrates notable advantages, our analysis reveals critical challenges and promising further directions, including (1) maintaining general capabilities, (2) exploring cost-effective frameworks, (3) overcoming generalization of TF alignment, (4) developing controllable TF alignment methods. This survey not only seeks to inspire further research interests in this area but also aims to guide the evolution of LLMs toward benefiting human society.

In summary, the main contributions of this paper are as follows:

- **First Review of TF Alignment**: To the best of our knowledge, this is the first comprehensive review of TF alignment methods for LLMs.
- **New Taxonomy Framework**: We present a systematic framework for TF alignment approaches, which organizes and categorizes the existing work in a structured way, covering pre-decoding interventions, in-decoding adjustments, and post-decoding refinement.
- **Prospective Research Roadmap**: We outline critical and challenging future directions for TF alignment, addressing key areas such as cost-effective framework, and controllable alignment. This roadmap provides a foundation for advancing TF alignment research towards more robust and inclusive LLMs.

2 Preliminary

This section systematically explores four pivotal dimensions of TF alignment, including alignment-related concepts, the rationale for choosing TF

alignment, scenarios where TF alignment is preferable, and methods to evaluate TF alignment approaches.

2.1 What are LLM Alignment and TF Alignment?

LLM alignment is the process of ensuring that language models behave in ways that align with human expectations, societal values, legal standards, and ethical principles. This review focuses on TF alignment, which encompasses functional alignment and normative alignment.

Functional alignment emphasizes the model's ability to perform tasks accurately and reliably. It can be further divided into two categories: (1) *Task-specific Alignment*: Tailoring the model for specific applications (e.g., Zhan et al. (2024)) to optimize the objective towards particular requirements. (2) *General Capability Alignment*: Ensuring the model maintains updated knowledge and strong generalization abilities across diverse tasks and scenarios (e.g., Zhang et al. (2024c)).

Normative alignment focuses on ethical, safety, and user-centric aspects. It also contains two main components: (1) *Public Safety and Ethical Alignment*: Ensuring model outputs comply with safety standards, ethical norms, and regulatory guidelines to avoid harmful content and data privacy violations (e.g., Fang et al. (2024)). The goal is to mitigate potential risks and prevent adverse impacts on society. (2) *Personalized Alignment*: Adapting the model's interaction style and outputs to individual preferences and usage patterns, balancing safety protocols with personalized experiences (e.g., Chen et al. (2024b)).

In conclusion, functional alignment ensures the model performs tasks effectively, while normative alignment guarantees that its operation remains safe and ethical.

2.2 Why Opt for TF Alignment?

Building upon the superficial nature of alignment tuning demonstrated by LIMA (Zhou et al., 2023), the extended empirical discovery in URIAL (Lin et al., 2023) demonstrates that TF alignment can match or even surpass FT-aligned LLM performance. TF alignment technology warrants deeper analysis to advance future LLM research. Therefore, we undertake a systematic investigation into TF alignment.

As listed in introduction, traditional alignment methods based on FT have one or more critical limitations in practice[1]. In contrast, TF alignment provides a versatile and efficient alternative. The core advantages of TF alignment are as follows.

Cost Efficiency and Environmental Sustainability (1) *No Training Overheads*: TF methods reduce computational costs (e.g., GPU hours) and hyperparameter tuning burdens since they eliminate the need to train LLMs. (2) *Reduced Data Dependency*: Unlike supervised fine-tuning (SFT), they often require no labeled datasets, minimizing annotation labor and data collection costs.

Operational Flexibility and Model Compatibility (1) *Plug-and-Play*: Techniques like in-context learning can be applied "on the fly" (Min et al., 2022), enabling rapid iteration (e.g., switching alignment objectives without retraining). (2) *Low Storage Overhead*: Unlike parameter-efficient fine-tuning (PEFT) methods (e.g., LoRA (Hu et al., 2021), TF alignment introduces low additional storage costs. (3) *Black-Box Adaptability*: Most TF alignment methods operate without accessing internal model parameters, making them compatible with both open- and closed-source models (e.g., Llama, GPT-4, Claude).

Slight Knowledge Degradation By avoiding parameter updates or making minor adjustments, TF alignment methods better preserve the original knowledge of LLMs.

2.3 How to Evaluate LLM Alignment?

The evaluation of LLM alignment effectiveness can be categorized into three primary paradigms based on assessment methods.

Benchmark-Driven Evaluation This quantitative approach utilizes standardized test suites with predefined metrics to measure alignment performance. For instance, MMLU (Hendrycks et al., 2020) and XSum (Narayan et al., 2018) benchmarks assess general capability retention after alignment, while perplexity measurements (Shannon, 1948) evaluate prediction quality by quantifying the cross-entropy between model outputs and reference distributions.

Human-Centric Evaluation This paradigm employs human judgment through two primary modalities: crowdsourcing annotation and expert analysis. Large-scale evaluations often leverage platforms like Amazon Mechanical Turk[2] to collect

[1]For example, modern adapter-based FT techniques can significantly reduce computational cost and are less energy-intensive, but the data dependency issue and the closed-source model accessibility issue still exist.

[2]https://www.mturk.com

judgments from workers. Also, domain-specific evaluations can be conducted by experts performing granular behavioral analyses.

Model-Assisted Evaluation Emerging automated methods employ auxiliary LLMs as evaluators through two principal strategies: prompt-based assessment and self-examination techniques. The former implements critic models (e.g., GPT-4) for direct quality scoring via instructional prompting, while the latter verifies alignment consistency through generated explanations that expose model reasoning patterns.

3 TF Alignment Methods

TF alignment methods can intervene at different stages of the generation pipeline without heavily retraining LLMs. Hence we categorize them into Pre-Decoding, In-Decoding, and Post-Decoding TF alignment, analyzing their application to unimodal LLMs and MLLMs. Figure 2 shows TF alignment methods at different stages, as well as their specific mechanisms.

3.1 Pre-Decoding TF Alignment

Pre-decoding TF alignment methods align models by modifying or detecting inputs before decoding without changing models' internal parameters. These approaches are lightweight and especially suitable for black-box LLMs.

Simple Prompt Engineering One of the simplest yet effective alignment strategies involves crafting specific prompts to elicit desired responses. ICL has emerged as a powerful tool for aligning LLMs with human preferences, a technique known as *In-Context Alignment (ICA)*. For instance, URIAL (Lin et al., 2023) demonstrates that just three in-context examples and a single system prompt can effectively align pre-trained LLMs, achieving performance comparable to FT methods while significantly reducing costs. Similarly, Anthropological Prompting (AlKhamissi et al., 2024) and VILMO (Duan et al., 2024) leverage specific prompts to enhance alignment. Besides, ICDPO (Song et al., 2025a) leverages ICL and instant scorer to enhance the final performance. To tackle the inflexibility of existing alignment methods when confronted with diverse social norms across different cultures and regions, CoSA (Zhang et al., 2024a) introduces safety configs, which allow models to dynamically adapt during inference based on these configurations, thereby catering to varying requirements.

Although MLLMs introduce a novel visual modality, enabling the embedding of harmful content within images to circumvent security mechanisms, the prompt engineering strategy in pre-decoding remains applicable. To defend against structured-based jailbreak attacks, AdaShield (Wang et al., 2024c) includes both manual static and adaptive defense prompts, the latter being iteratively optimized through collaboration between the target MLLM and an LLM-based defense prompt generator. Similarly, to address the limitation that static safety guidelines fail to account for specific risks in different multimodal contexts, ICD (Wei et al., 2024), RapGuard (Jiang et al., 2024), and BBA (Zhao et al., 2024) generate scenario-specific security prompts. They leverage multimodal reasoning to dynamically identify and mitigate risks.

Prompts can also act as self-reminders (Xie et al., 2023), packaging user queries and prompting ChatGPT to respond responsibly. SKIG (Sel et al., 2024) designs multi-dimensional prompts that stimulate the model's sense of responsibility, exploring the consequences of decisions from multiple stakeholders' perspectives. This approach enhances the model's moral reasoning ability.

Enhanced Prompt Strategy Beyond simple prompts, this category encompasses methods that enhance prompts through additional techniques or iterative refinement. Certain human values to be aligned often vary with time and place, OPO (Xu et al., 2023) uses retrieval-augmented-generation (RAG) can address the ever-changing nature of alignment in human values. AUTOCAP (Zhang et al., 2024d) integrates Chain-of-Thought (CoT) reasoning paths across languages to improve cross-lingual alignment (Shi et al., 2022; Tanwar et al., 2023; Qin et al., 2023). P-Aligner (Song et al., 2025b) automatically rewrites user instructions with pre-defined principles, significantly boosting downstream LLM alignment. Other methods focus on feeding inputs in a more structured way. For example, Zhang et al. (2024e) introduce goal priority through few-shot prompting, instructing the LLM to prioritize safety over helpfulness. PRETTY (Zhan et al., 2024) improves alignment by adding task-related prior markers to the input prefix, narrowing the performance gap between TF models and FT models. BPO (Cheng et al., 2024) op-

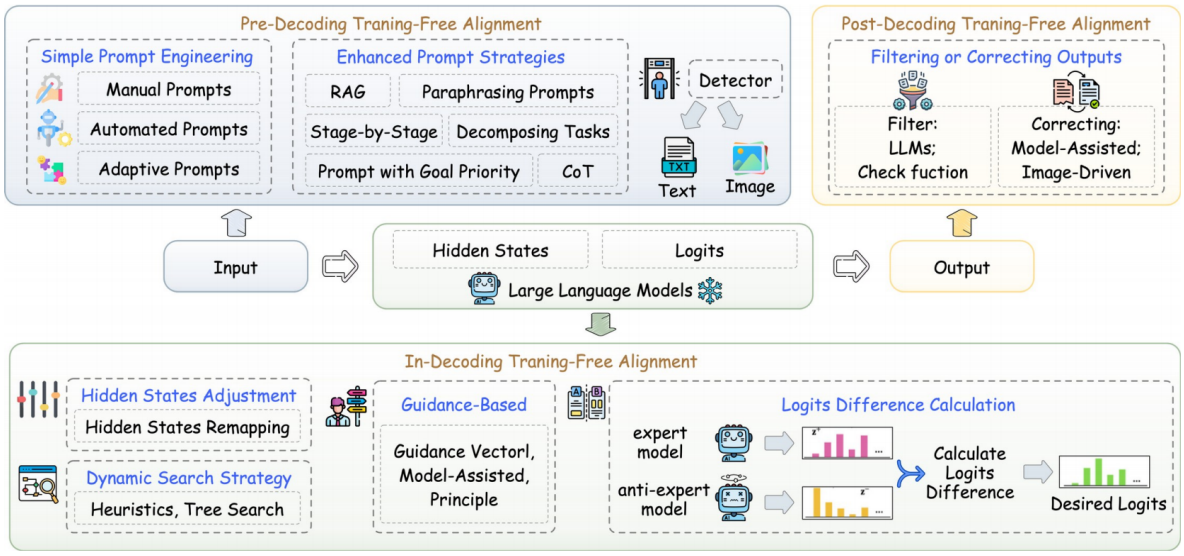

Figure 2: A conceptual framework illustrating training-free (TF) alignment strategies for large language models (LLMs), categorized into pre-decoding, in-decoding, and post-decoding stages.

timizes user prompts to suit LLMs' input understanding, ensuring user intents are realized without updating internal parameters. PEARL (Fu et al., 2024a) paraphrases questions in expressions preferred by the model, ensuring better alignment with user expectations.

Additionally, the decomposition of complex problems into subproblems has proven effective. For example, MIXALIGN (Zhang et al., 2024c) and DUAT (Huang et al., 2024d) break down difficult problems into subtasks and designs corresponding prompts step-by-step, fully utilizing the generation and reasoning abilities of LLMs to dynamically solve knowledge alignment challenges. Another way uses prompts at the first stage, followed by additional operations to further refine the output. PICA (Liu et al., 2024d) operates in two stages: first, the model generates prior response tokens via ICL while extracting an ICL vector; second, this ICL vector guides the model to generate desired responses without requiring further demonstrations. This approach ensures consistent alignment while minimizing the need for extensive prompting.

Detector-Based VLMGUARD (Du et al., 2024) utilizes unlabeled user prompt data for malicious prompts detection, effectively distinguishing between malicious and benign samples without additional manual annotation. CIDER (Xu et al., 2024a) detects malicious image inputs by analyzing the semantic similarity between text and image modalities. Token Highlighter (Hu et al., 2024) locates the jailbreak-critical tokens and use soft removal technique to align. To reduce the cost (e.g., substantial memory requirements and latency) of detector with billions of parameters, HarmAug (Lee et al., 2024) distills a large teacher safety guard model into a smaller one.

Key Insights Pre-decoding methods aim to reformulate queries, append exemplars, or leverage detectors before feeding inputs into LLMs, ensuring model outputs align with human standards. These approaches offer advantages of simplicity and broad applicability, working seamlessly with both open-source and closed-source models. However, they suffer from generalization limitations due to reliance on few-shot examples or manual prompt engineering. For instance, CoSA (Zhang et al., 2024a) introduces safety configurations to enable dynamic inferential adaptation, yet this only addresses safety scenarios—cross-scenario generalization, cultural variance, and model-agnostic adaptability remain underexplored. Additionally, most prompt engineering techniques are constrained by context window limits (e.g., LLaMA-2-13B's 4096-token boundary (Machlab and Battle, 2024)) or suffer semantic drift (Ji et al., 2023) due to excessively long inputs. The subsequent in-decoding and post-decoding methods tackle these limitations from alternative perspectives.

3.2 In-Decoding TF Alignment

Aligning LLMs before decoding often appears ideal following the principle of "an ounce of prevention is worth a pound of cure". However, not all undesired behavior can be fully addressed at this stage. This motivates in-decoding TF alignment, which adjusts model behaviors during decoding. These methods typically modify hidden states or logits, through remapping hidden states, calculating logits difference, exacting guidance signals, or employing search strategies.

Hidden States Adjustment The disjoint training protocols between visual and linguistic modules in VLMs induce latent space fragmentation, where their representations form distinct cluster regions with significant distributional discrepancy, manifesting diminished safety-aligned capabilities. To address these issues, CMRM (Liu et al., 2024b) and VLM-Guard (Liu et al., 2025) mitigate this by pulling hidden states back into the representation space optimized by LLMs, thus restoring secure alignment capabilities.

Logits Difference Calculation Several methods modify the logits of tokens during generation to align outputs with desired behaviors. (Liu et al., 2024c,a; Shi et al., 2024; Fang et al., 2024; Zhu et al., 2025b) combine the logits of the alignment model and a reference model to guide generation, achieving FT-like effects without parameter updates. δ -UNLEARNING (Huang et al., 2024b) learns an offset from logits comparisons between smaller models, applying it to larger black-box LLMs to adjust their predictive behavior. Additionally, CoCA (Gao et al., 2024a) calibrates the output distribution by amplifying the model's response to safety prompts. The method calculates logits difference before and after safety prompt insertion, ensuring robust alignment. Recently, IVG (Liu et al., 2024e) uses implicit and explicit value functions to guide language model decoding at token and chunk-level respectively, efficiently aligning LLMs purely at inference time. Contrastive decoding has also been adopted. For example, MCA (Fu et al., 2024b) constructs expert and adversarial prompts to promote or suppress aligned targets dynamically. Lately, Linear Alignment (Gao et al., 2024b) relies on a novel parameterization for policy optimization under divergence constraints and estimates the preference direction using self-contrastive decoding.

Guidance-Based For guidance-based approaches, (Qian et al., 2024; Wang et al., 2024b; Huang et al., 2024c; Bhattacharjee et al., 2024; Cao et al., 2025; Li et al., 2025b) extract guidance vectors to adjust the activation states of the target model during inference, achieving alignment. To predict next-token rewards for efficient and effective autoregressive generation, GenARM (Xu et al., 2025) and CARDS (Li et al., 2025a) leverage the reward model to guide frozen LLMs toward expected distribution. PAD (Chen et al., 2024b) and ARGS (Khanov et al., 2024) leverage token-level rewards to guide the decoding process, dynamically guiding the base model's predictions. Nudging (Fei et al., 2024) and WSD (Song et al., 2025c) combine a large base model with a much smaller aligned model. CAVGAN (Li et al., 2025c) utilizes generative adversarial network to learn the security judgment boundary inside the LLM to achieve efficient alignment. RDS (Zeng et al., 2025) designs a root classifier based on the discriminative capacity of queries, and then reorders the token and prioritizes the benign token. To directly align model outputs with human preferences, OPAD (Zhu et al., 2025a) guides the generation of a final output that aligns with the target principle in a token-by-token manner.

Dynamic Search Strategy Search strategies have also been frequently used for alignment. RAIN (Li et al., 2024b) mirrors human behavioral patterns, which allows LLMs to evaluate their own generation and use the evaluation results to guide rewind and generation for self-alignment. Similarly, DeAL (Huang et al., 2024a) views decoding as a heuristic-guided search process and facilitates the use of a wide variety of alignment objectives. TreeBoN (Qiu et al., 2024) combines tree search strategies with Best-of-N (BoN) sampling, iteratively expanding high-quality partial responses while pruning low-quality candidates to improve alignment quality and efficiency. DARWIN (Hung et al., 2024) achieves the right balance between exploration and exploitation of rewards during decoding with evolutionary heuristics. Additionally, Zhu et al. (2024) proposed PAS, which includes direction search and distance search. They focuse on the alignment with personal traits and develop an activation intervention optimization method to enhance LLMs' ability to efficiently align with individual behavioral preferences using minimal data and computational resources.

Key Insights In-decoding TF alignment methods enable dynamic, fine-grained output control via hidden state adjustment, logit modification, tree search strategies, etc. For MLLMs, projecting hidden states into LLM-optimized representation spaces shows promise in resolving multimodal integration challenges. While highly effective for alignment and alleviating certain generalization issues, this paradigm often requires access to model internals, restricting applicability in black-box systems. Moreover, techniques like Trustworthiness and Category-Specific Steering (Qian et al., 2024; Bhattacharjee et al., 2024), which modify hidden or activation states, can inadvertently compromise the model's pre-existing task-specific knowledge. These trade-offs directly inform the research directions outlined in Section 4.2.

3.3 Post-Decoding TF Alignment

In-decoding TF alignment methods often require access to token logits and vocabulary, restricting their applicability to models from the same series. To overcome these limitations, post-decoding TF alignment methods are proposed, which operate on the generated outputs without accessing the model internals.

Post-decoding TF alignment methods often adopt strategies to filter or correct outputs. Phute et al. (2024) utilize inherent language comprehension of LLMs to self-examine generated text. RA-LLM (Cao et al., 2024) uses an alignment check function, enhancing robustness by randomly discarding portions of the input request to mitigate adversarial perturbations. For correction-based approaches, Aligner (Ji et al., 2024a) trains a separate model to learn the residual between the initial and aligned outputs. This plug-and-play method redistributes initial answers into more helpful and harmless responses.

Multimodal applications introduce additional challenges, such as visual inputs that may contain toxicity. ETA (Ding et al., 2024) uses a multimodal evaluator to assess visual inputs with the CLIP score. If need to align, ETA can operate shallow alignment (interference prefix) and deep alignment (sentence-level best-of-N searching). ECSO (Gou et al., 2024) also reviews responses first. If deemed unsafe, it converts images into text via query-aware transformation, enabling the pre-aligned LLM to handle both text-based and image-embedded malicious content for safe output generation. Chen et al. (2023) propose an instruction based on self-moderation, deciding whether to respond based on privacy concerns. DPS (Zhou et al., 2024b) integrates LLM security checkers to filter responses, defending against visual attacks while maintaining clean input performance. This reduces harmful content while preserving model performance.

Key Insights Analogous to pre-decoding approaches, post-decoding TF alignment methods are model agnostic. Also, they mitigate the generalization constraints seen in pre-decoding stages, offering heightened versatility. In MLLM contexts, integrating visual-textual data enables dynamic defense mechanisms, though this amplifies alignment complexity. A notable trade-off is the latency introduced by filtering or correction processes.

3.4 Quantitative Analysis

To have an insight into diverse alignment methods, we conduct experiments on llama2-7b-chat with safety alignment for illustration.

Datasets Following SCANS (Cao et al., 2025), we select AdvBench and TruthfulQA (Lin et al., 2021) datasets to evaluate the safety and helpfulness of alignment methods. Building on this, we incorporate a more challenging dataset, SafeEdit (Wang et al., 2024a), to assess how the aligned models perform against adversarial attacks.

Methods For FT alignment methods, we select SafeDecoding (Xu et al., 2024b), a relatively new method that accounts for jailbreak attacks and demonstrates superior performance. For TF alignment methods, we select one recent approach for each of three stages (Lin et al., 2023; Cao et al., 2025, 2024).

Results The results are shown in Table 1. Notably, TF alignment methods can match or even exceed the safety and helpfulness performance of FT alignment methods. For example, SCANS outperforms SafeDecoding on SafeEdit and TruthfulQA datasets. It also shows performance comparable to SafeDecoding on AdvBench. This reflects the effectiveness of TF alignment as a supplement to FT alignment, demonstrating that TF methods merit deeper investigation, particularly in scenarios where FT alignment faces practical constraints (e.g., incompatibility with closed-source models, significant knowledge degradation during fine-tuning). However, TF alignment methods also exhibit certain limitations. For example, TF methods like URIAL and RA-LLM compromise the model's helpfulness, which certainly damages the

Model	Type	AdvBench ↑	SafeEdit ↑	TruthfulQA ↓
llama2-7b-chat	Defaults	99.78	37.60	5.05
SafeDecoding	FT Alignment	100.00	94.60	54.44
URIAL	Pre-Decoding TF Alignment	99.34	66.60	15.94
SCANS	In-Decoding TF Alignment	99.34	97.80	0.80
RA-LLM	Post-Decoding TF Alignment	100.00	98.00	36.12

Table 1: The numerical comparison of the performance of llama2-7b-chat with different safety alignment methods when facing with malicious questions, adversarial attacks and safe questions. On AdvBench and SafeEdit datasets, numerical values denote the Defense Success Rate, where ↑ indicates that higher values are preferable. On TruthfulQA, the numerical values represent the Benign Refusing Rate, and ↓ signifies that lower values are more desirable.

helpful performance on TruthfulQA, albeit such damage is less severe compared to the FT method SafeDecoding. More challenges will be described in the next section. We also analyze practical deployment considerations for different TF alignment methods in Table 2 of Appendix A.

Please refer to Appendix A for more details on datasets, evaluation criteria, result analysis, and comparison of different TF alignment methods.

4 Discussion

While the above TF alignment methods demonstrate notable advantages, significant challenges still exist. Meanwhile, there are many opportunities to make TF alignment more effective, efficient, and adaptive.

4.1 Open Challenges

While TF alignment offers practical advantages, our systematic analysis reveals three fundamental challenges that constrain its broader adoption: (1) *Degradation of General Capabilities*: While most TF alignment methods preserve model knowledge integrity better than FT alignment methods, certain in-decoding methods (Qian et al., 2024; Bhattacharjee et al., 2024) potentially compromise general performance. Because these methods adjust the hidden states or activation states of the target model during inference, such operations undermine the model's inherent knowledge for other tasks. This interference disrupts the pre-trained model's generalized ability to handle diverse tasks, as the adjusted states may no longer align with the optimal representations learned for cross-task scenarios. Consequently, the model's performance in non-target tasks might deteriorate, highlighting the trade-off between task-specific adaptation and maintenance of general capabilities in alignment strategies. (2) *Increased Inference Overhead*: TF alignment approaches often introduce additional latency. Whether through input detection, logits adjustment, or modification of generated responses, the additional computational steps result in higher latency. (3) *Difficult to Generalize*: TF alignment cannot leverage extensive alignment data. As a result, it may fail to generalize effectively to unseen or more challenging scenarios (e.g. adversarial attacks), as it lacks exposure to diverse or complex cases.

4.2 Future Directions

Maintain General Capabilities Develop mechanisms to maintain or enhance the general capabilities of LLMs while performing alignment, such as incorporating auxiliary loss functions that balance alignment objectives with model performance. Another promising way is exploring lightweight interventions in hidden states or logits that minimize disruption to the model's original outputs, leveraging techniques like sparse interventions. For instance, when adjusting hidden states, task-relevant neurons (Leng and Xiong, 2025; Tang et al., 2024) can be identified, eliminating the need to modify the entire state space and thereby preserving a significant portion of the original performance.

Explore Unified Metrics and Cost-Effective Frameworks To compare various alignment methods more thoroughly, unified computational overhead metrics to evaluate memory usage, CPU/GPU utilization, and time taken for alignment steps need to be established. Additionally, cost-effective algorithms for input detection, logits adjustment, and response modification to reduce latency should be designed. For instance, HarmAug (Lee et al., 2024) distills a large teacher safety guard model into a smaller one. Also, it is necessary to investigate on-the-fly optimization strategies (Zhu et al., 2025a) that dynamically adapt computational resources based on complexity of

tasks or efficiency requirements of users.

Overcome Insufficient Generalization When addressing the diverse scenarios of different cultures, applications, or users, it is essential to establish corresponding principles for guidance. For instance, CoSA (Zhang et al., 2024a) introduces safety configs, which allow models to dynamically adapt during inference based on these configurations, thereby catering to varying requirements. However, this only accounts for different safety scenarios, while the generalization across other scenarios and diverse cultures, as well as the adaptability of various models, remains equally worthy of exploration. Additionally, incorporating counterfactual samples into prompts can break spurious correlations, compelling the model to learn causal features and thereby enhancing its generalization capability. Meta-learning (Finn et al., 2017; Khoee et al., 2024) techniques can also be leveraged, enabling TF alignment methods to quickly adapt to new, unseen scenarios with minimal additional data. Furthermore, explore hybrid approaches that combine the strengths of FT (rich alignment data) and TF alignment (no parameter updates) to achieve better generalization.

TF Alignment for Uni-Modal Model Although some efforts have been made in the field of TF alignment for MLLMs, they are currently limited to scenarios where the input is multimodal (e.g., image and text) and the output is purely textual. Extending alignment to cases where the output includes multimodal content, such as images, introduces new challenges and complexities (Xiong et al., 2024; Team, 2024; Xie et al., 2024; Zhou et al., 2024a), making it a promising and worthwhile direction for future research. Ji et al. (2024b) have made initial strides in aligning all-modality models with human intentions through FT methods. However, research on TF alignment for such Uni-Modal Models is still lacking.

Controllable TF Alignment Interpretability can help address this problem (Wu et al., 2024). It locates and edits fine-grained features, which can be used to align LLMs to human values. For example, by feeding prompts with negative and positive prefixes into LLMs, Leong et al. (2023) analyze internal contextualized representations to identify the toxicity direction of each attention head. However, many interpretability studies are primarily based on toy or theoretical models. Developing more practical interpretability TF alignment methods is necessary. Another way is to use principle-based guidance. OPAD (Zhu et al., 2025a) designs a principle-guided reward function to align model outputs with human preferences without FT. Moreover, considering the advantages of TF alignment in customization, explore the use of controllable TF alignment for personalized scenarios (Zhu et al., 2025a; Chen et al., 2024b).

5 Conclusion

In this paper, we present a comprehensive survey of training-free (TF) alignment methods for LLMs, making three significant contributions to the field. First, we establish a conceptual framework for the TF alignment. Second, we propose a novel taxonomy that categorizes TF alignment techniques into three distinct paradigms: pre-decoding interventions, in-decoding adjustments, and post-decoding refinement strategies. Finally, we identify promising research directions, emphasizing the need for effective, efficient, and adaptive TF alignment methods.

This survey is expected to serve both as an introductory guide for newcomers and a reference manual for practitioners, aiming to accelerate progress in developing efficient, ethical, and human-aligned LLM systems. We hope that our investigation will inspire novel research addressing the critical challenges and future directions of TF alignment.

Limitations

While this survey provides a systematic overview of TF alignment methods for LLMs, several inherent limitations in its scope need more discussion: Firstly, in this review, base models must be sufficiently capable of responding effectively to TF alignment techniques, which means that these models might require prior fine-tuning. Secondly, we pay more attention to safety TF alignment due to page limitation, which makes it infeasible to provide a detailed overview of all TF alignment approaches. Thus, we choose to focus intensively on one category. Finally, experiments just conducted on llama2-7b-chat because the nature of our literature survey. It is hoped that our preliminary exploration will inspire further researches on TF alignment.

Ethics Statement

Our work is entirely at the methodological level, which means that there will not be any negative social impacts.

Acknowledgments

This work was supported by a grant from the National Natural Science Foundation of China (NSFC) project (No. 62276193), the Key Laboratory of Computing Power Network and Information Security, Ministry of Education under Grant No. 2024ZD027, and the Fundamental Research Funds for the Central Universities, China (No. 2042022dx0001).

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Michael Beukman, Devon Jarvis, Richard Klein, Steven James, and Benjamin Rosman. 2023. Dynamics generalisation in reinforcement learning via adaptive context-aware policies. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Amrita Bhattacharjee, Shaona Ghosh, Traian Rebedea, and Christopher Parisien. 2024. Towards inference-time category-wise safety steering for large language models. *Preprint*, arXiv:2410.01174.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2024. Defending against alignment-breaking attacks via robustly aligned LLM. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10542–10560, Bangkok, Thailand. Association for Computational Linguistics.

Zouying Cao, Yifei Yang, and Hai Zhao. 2025. Scans: Mitigating the exaggerated safety for llms via safety-conscious activation steering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23523–23531.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Meiqi Chen, Yubo Ma, Kaitao Song, Yixin Cao, Yan Zhang, and Dongsheng Li. 2024a. Improving large language models in event relation logical prediction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9451–9478, Bangkok, Thailand. Association for Computational Linguistics.

Ruizhe Chen, Xiaotian Zhang, Meng Luo, Wenhao Chai, and Zuozhu Liu. 2024b. Pad: Personalized alignment of llms at decoding-time. *Preprint*, arXiv:2410.04070.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023. Can language models be instructed to protect personal information? *Preprint*, arXiv:2310.02224.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2024. 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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yi Ding, Bolian Li, and Ruqi Zhang. 2024. Eta: Evaluating then aligning safety of vision language models at inference time. *Preprint*, arXiv:2410.06625.

Xuefeng Du, Reshmi Ghosh, Robert Sim, Ahmed Salem, Vitor Carvalho, Emily Lawton, Yixuan Li, and Jack W. Stokes. 2024. Vlmguard: Defending vlms against malicious prompts via unlabeled data. *Preprint*, arXiv:2410.00296.

Shitong Duan, Xiaoyuan Yi, Peng Zhang, Tun Lu, Xing Xie, and Ning Gu. 2024. DENEVIL: TOWARDS DECIPHERING AND NAVIGATING THE ETHICAL VALUES OF LARGE LANGUAGE MODELS VIA INSTRUCTION LEARNING. In *The Twelfth*

International Conference on Learning Representations.

Hainan Fang, Di Huang, Yuanbo Wen, Yunpu Zhao, Tonghui He, QiCheng Wang, Shuo Wang, Rui Zhang, and Qi Guo. 2024. GOOD: Decoding-time black-box LLM alignment.

Yu Fei, Yasaman Razeghi, and Sameer Singh. 2024. Nudging: Inference-time alignment via model collaboration. *Preprint*, arXiv:2410.09300.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR.

Junbo Fu, Guoshuai Zhao, Yimin Deng, Yunqi Mi, and Xueming Qian. 2024a. Learning to paraphrase for alignment with LLM preference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2394–2407, Miami, Florida, USA. Association for Computational Linguistics.

Tingchen Fu, Yupeng Hou, Julian McAuley, and Rui Yan. 2024b. Unlocking decoding-time controllability: Gradient-free multi-objective alignment with contrastive prompts. *Preprint*, arXiv:2408.05094.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

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

Songyang Gao, Qiming Ge, Wei Shen, Shihan Dou, Junjie Ye, Xiao Wang, Rui Zheng, Yicheng Zou, Zhi Chen, Hang Yan, Qi Zhang, and Dahua Lin. 2024b. Linear alignment: a closed-form solution for aligning human preferences without tuning and feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XVII*, page 388–404, Berlin, Heidelberg. Springer-Verlag.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. 2024. Token highlighter: Inspecting and mitigating jailbreak prompts for large language models. In *AAAI Conference on Artificial Intelligence*.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. 2024a. Deal: Decoding-time alignment for large language models. *Preprint*, arXiv:2402.06147.

James Y. Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024b. Offset unlearning for large language models. *Preprint*, arXiv:2404.11045.

Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. 2024c. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10943–10959, Bangkok, Thailand. Association for Computational Linguistics.

Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, and Bing Qin. 2024d. Aligning translation-specific understanding to general understanding in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5028–5041, Miami, Florida, USA. Association for Computational Linguistics.

Chia-Yu Hung, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2024. Inference time alignment with reward-guided tree search. *Preprint*, arXiv:2406.15193.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. 2024a. Aligner: Efficient alignment by learning to correct. *Preprint*, arXiv:2402.02416.

Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, Mohan Wang, Josef Dai, Tianyi Qiu, Hua Xu, Dong Li, Weipeng Chen, Jun Song, Bo Zheng, and Yaodong Yang. 2024b. Align anything: Training all-modality models to follow instructions with language feedback. *Preprint*, arXiv:2412.15838.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Yilei Jiang, Yingshui Tan, and Xiangyu Yue. 2024. Rapguard: Safeguarding multimodal large language models via rationale-aware defensive prompting. *Preprint*, arXiv:2412.18826.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. 2024. Args: Alignment as reward-guided search. *ArXiv*, abs/2402.01694.

Arsham Gholamzadeh Khoee, Yinan Yu, and Robert Feldt. 2024. Domain generalization through meta-learning: A survey. *Artificial Intelligence Review*, 57(10):285.

Seanie Lee, Haebin Seong, Dong Bok Lee, Minki Kang, Xiaoyin Chen, Dominik Wagner, Yoshua Bengio, Juho Lee, and Sung Ju Hwang. 2024. Harmaug: Effective data augmentation for knowledge distillation of safety guard models. *Preprint*, arXiv:2410.01524.

Yongqi Leng and Deyi Xiong. 2025. Towards understanding multi-task learning (generalization) of llms via detecting and exploring task-specific neurons. *Preprint*, arXiv:2407.06488.

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. Self-detoxifying language models via toxification reversal. *Preprint*, arXiv:2310.09573.

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

Qing Li, Jiahui Geng, Derui Zhu, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. 2025b. Internal activation revision: Safeguarding vision language models without parameter update. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27428–27436.

Xiaohu Li, Yunfeng Ning, Zepeng Bao, Mayi Xu, Jianhao Chen, and Tieyun Qian. 2025c. CAVGAN: Unifying jailbreak and defense of LLMs via generative adversarial attacks on their internal representations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6664–6678, Vienna, Austria. Association for Computational Linguistics.

Yichuan Li, Kaize Ding, Jianling Wang, and Kyumin Lee. 2024a. Empowering large language models for textual data augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12734–12751, Bangkok, Thailand. Association for Computational Linguistics.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2024b. RAIN: Your language models can align themselves without finetuning. In *The Twelfth International Conference on Learning Representations*.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Alisa Liu, Xiaochuang Han, Yizhong Wang, Yulia Tsvetkov, Yejin Choi, and Noah A. Smith. 2024a. Tuning language models by proxy. *Preprint*, arXiv:2401.08565.

Qin Liu, Chao Shang, Ling Liu, Nikolaos Pappas, Jie Ma, Neha Anna John, Srikanth Doss, Lluis Marquez, Miguel Ballesteros, and Yassine Benajiba. 2024b. Unraveling and mitigating safety alignment degradation of vision-language models. *Preprint*, arXiv:2410.09047.

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2025. Vlm-guard: Safeguarding vision-language models via fulfilling safety alignment gap. *Preprint*, arXiv:2502.10486.

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. *Preprint*, arXiv:2402.02992.

Zhenyu Liu, Dongfang Li, Xinshuo Hu, Xinping Zhao, Yibin Chen, Baotian Hu, and Min Zhang. 2024d. Take off the training wheels! progressive in-context learning for effective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2743–2757, Miami, Florida, USA. Association for Computational Linguistics.

Zhixuan Liu, Zhanhui Zhou, Yuanfu Wang, Chao Yang, and Yu Qiao. 2024e. Inference-time language model alignment via integrated value guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4181–4195, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Machlab and Rick Battle. 2024. Llm in-context recall is prompt dependent. *Preprint*, arXiv:2404.08865.

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.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

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.

Mansi Phute, Alec Helbling, Matthew Hull, ShengYun Peng, Sebastian Szyller, Cory Cornelius, and Duen Horng Chau. 2024. Llm self defense: By self examination, llms know they are being tricked. *Preprint*, arXiv:2308.07308.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. 2024. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4864–4888, Bangkok, Thailand. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. 2024. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *Preprint*, arXiv:2410.16033.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *ArXiv*, abs/2310.20081.

Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Bilgehan Sel, Priya Shanmugasundaram, Mohammad Kachuee, Kun Zhou, Ruoxi Jia, and Ming Jin. 2024. Skin-in-the-game: Decision making via multi-stakeholder alignment in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13921–13959, Bangkok, Thailand. Association for Computational Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *Preprint*, arXiv:2309.15025.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Ruizhe Shi, Yifang Chen, Yushi Hu, Alisa Liu, Hannaneh Hajishirzi, Noah A Smith, and Simon S Du. 2024. Decoding-time language model alignment with multiple objectives. *arXiv preprint arXiv:2406.18853*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology? *Preprint*, arXiv:2306.03809.

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.

Feifan Song, Shaohang Wei, Wen Luo, Yuxuan Fan, Tianyu Liu, Guoyin Wang, and Houfeng Wang. 2025c. Well begun is half done: Low-resource preference alignment by weak-to-strong decoding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12654–12670, Vienna, Austria. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *ArXiv*, abs/2305.05940.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024a. Detoxifying large language models via knowledge editing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3093–3118, Bangkok, Thailand. Association for Computational Linguistics.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Mozhi Zhang, Ke Ren, Botian Jiang, and Xipeng Qiu. 2024b. InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10460–10479, Miami, Florida, USA. Association for Computational Linguistics.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. 2024c. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *Preprint*, arXiv:2403.09513.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Zixu, Zhu, Xiang-Bo Mao, Sitaram Asur, Na, and Cheng. 2024d. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *Preprint*, arXiv:2407.16216.

Zeming Wei, Yifei Wang, Ang Li, Yichuan Mo, and Yisen Wang. 2024. Jailbreak and guard aligned language models with only few in-context demonstrations. *Preprint*, arXiv:2310.06387.

Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, and Ninghao Liu. 2024. Usable xai: 10 strategies towards exploiting explainability in the llm era. *Preprint*, arXiv:2403.08946.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024. Show-o: One single transformer to unify multimodal understanding and generation. *Preprint*, arXiv:2408.12528.

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.

Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, Chaofan Tao, Shen Yan, Huaxiu Yao, Lingpeng Kong, Hongxia Yang, Mi Zhang, Guillermo Sapiro, Jiebo Luo, Ping Luo, and Ngai Wong. 2024. Autoregressive models in vision: A survey. *Preprint*, arXiv:2411.05902.

Chunpu Xu, Steffi Chern, Ethan Chern, Ge Zhang, Zekun Wang, Ruibo Liu, Jing Li, Jie Fu, and Pengfei Liu. 2023. Align on the fly: Adapting chatbot behavior to established norms. *Preprint*, arXiv:2312.15907.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. 2025. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *Preprint*, arXiv:2410.08193.

Yue Xu, Xiuyuan Qi, Zhan Qin, and Wenjie Wang. 2024a. Cross-modality information check for detecting jailbreaking in multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13715–13726, Miami, Florida, USA. Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Niu Luyao, Jia Jinyuan, Bill Y. Lin, and Radha Poovendran. 2024b. Safedecoding: Defending against jailbreak attacks via safety-aware decoding. In *Annual Meeting of the Association for Computational Linguistics*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Xinyi Zeng, Yuying Shang, Jiawei Chen, Jingyuan Zhang, and Yu Tian. 2025. Root defence strategies: Ensuring safety of llm at the decoding level. *Preprint*, arXiv:2410.06809.

Runzhe Zhan, Xinyi Yang, Derek Wong, Lidia Chao, and Yue Zhang. 2024. Prefix text as a yarn: Eliciting non-English alignment in foundation language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12131–12145, Bangkok, Thailand. Association for Computational Linguistics.

Jingyu Zhang, Ahmed Elgohary, Ahmed Magooda, Daniel Khashabi, and Benjamin Van Durme. 2024a. Controllable safety alignment: Inference-time adaptation to diverse safety requirements. *Preprint*, arXiv:2410.08968.

LingXi Zhang, Yue Yu, Kuan Wang, and Chao Zhang. 2024b. ARL2: Aligning retrievers with black-box large language models via self-guided adaptive relevance labeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3708–3719, Bangkok, Thailand. Association for Computational Linguistics.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2024c. The knowledge alignment problem: Bridging human and external knowledge for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2025–2038, Bangkok, Thailand. Association for Computational Linguistics.

Yongheng Zhang, Qiguang Chen, Min Li, Wanxiang Che, and Libo Qin. 2024d. AutoCAP: Towards automatic cross-lingual alignment planning for zero-shot chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9191–9200, Bangkok, Thailand. Association for Computational Linguistics.

Zhexin Zhang, Junxiao Yang, Pei Ke, Fei Mi, Hongning Wang, and Minlie Huang. 2024e. Defending large language models against jailbreaking attacks through goal prioritization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8865–8887, Bangkok, Thailand. Association for Computational Linguistics.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. 2023. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311, Singapore. Association for Computational Linguistics.

Xueliang Zhao, Xinting Huang, Tingchen Fu, Qintong Li, Shansan Gong, Lemao Liu, Wei Bi, and Lingpeng Kong. 2024. BBA: Bi-modal behavioral alignment for reasoning with large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7255–7279, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024a. Transfusion: Predict the next token and diffuse images with one multi-modal model. *Preprint*, arXiv:2408.11039.

Qi Zhou, Tianlin Li, Qing Guo, Dongxia Wang, Yun Lin, Yang Liu, and Jin Song Dong. 2024b. Defending lvlms against vision attacks through partial-perception supervision. *Preprint*, arXiv:2412.12722.

Mingye Zhu, Yi Liu, Lei Zhang, Junbo Guo, and Zhendong Mao. 2025a. On-the-fly preference alignment via principle-guided decoding. In *The Thirteenth International Conference on Learning Representations*.

Minjun Zhu, Linyi Yang, and Yue Zhang. 2024. Personality alignment of large language models. *Preprint*, arXiv:2408.11779.

Wenhong Zhu, Ruobing Xie, Weinan Zhang, and Rui Wang. 2025b. Flexible realignment of language models. *Preprint*, arXiv:2506.12704.

Yuchen Zhuang, Haotian Sun, Yue Yu, Rushi Qiang, Qifan Wang, Chao Zhang, and Bo Dai. 2024. Hydra: Model factorization framework for black-box llm personalization. *arXiv preprint arXiv:2406.02888*.

A Appendix

Taking safety alignment as an example, the analysis in Table 1 compares the safety alignment performance of different methods applied to llama2-7b-chat. Detailed information regarding datasets, evaluation criteria, results analysis, and comparisons of various TF alignment methods is presented as follows.

A.1 Dataset and Evaluation Criteria

Following SCANS (Cao et al., 2025), we selected AdvBench and TruthfulQA (Lin et al., 2021) datasets to evaluate the safety and helpfulness of aligned models. Building on this, we incorporated the more challenging SafeEdit (Wang et al., 2024a) dataset to assess how the aligned models perform against adversarial attacks. AdvBench consists of 456 harmful queries. SafeEdit, comprising 500 adversarial attacks across 9 unsafe scenarios, is used to evaluate the robustness of these alignment methods under extreme adversarial conditions. Besides, TruthfulQA is used as the test set for evaluating helpfulness, consisting of 753 benign questions.

We comprehensively assess the effectiveness of different alignment methods from two crucial dimensions: the Defense Success Rate, which reflects safety, and the Benign Refusing Rate, which reflects helpfulness. Regarding the Defense Success Rate, we employ the safety judgement classifier proposed by Wang et al. (2024a) to evaluate whether a response is safe. When evaluating the Benign Refusing Rate, we utilize the classic keyword detection method. If responses to safety questions following the application of an alignment method contain more refusal-related keywords, the method demonstrates lower utility.

A.2 More Numerical Analysis

Data from Table 1 reveals that TF alignment methods can rival or even surpass the safety and helpfulness performance of FT alignment methods. For example, while SCANS demonstrates safety performance comparable to that of FT-based SafeDecoding, it exhibits superior usefulness. Furthermore, all TF alignment methods presented in the table demonstrate higher levels of helpfulness. These findings underscore that TF alignment methods merit further investigation and refinement, particularly in contexts where FT alignment methods face constraints, such as severe knowledge distortion and model accessibility.

Specifically, URIAL demonstrates poor performance on the SafeEdit dataset because the method uses only three examples as prompts, which are fed into the LLM alongside the question. This approach fails to generalize well to more challenging adversarial attack problems, as discussed in Section 4.1. Additionally, comparing the performance of different methods on TruthfulQA can partially reflect the degree of impairment to the model's inherent capabilities. It is evident that FT alignment methods cause the most severe knowledge impairment, while TF alignment methods exhibit a milder impact by comparison. For further comparison, experiments can be conducted on datasets such as MMLU and XSum (as discussed in Section 2.3).

A.3 When and How to Choose TF Alignment?

We conclude that TF alignment methods are particularly well-suited for the following scenarios: (1) *Resource-Constrained Settings*: When computational resources or training data are limited, TF alignment methods offer a lightweight and cost-effective solution. (2) *Black-Box Models*: When the model parameters are inaccessible (e.g., closed-source or proprietary models), TF alignment methods provide a viable alternative to FT. (3) *Rapid Deployment*: When quick adaptation is required, TF alignment methods like in-context learning or decoding-time adjustments enable immediate alignment without lengthy FT processes. (4) *Preserving Knowledge*: When FT alignment risks the degrading the model's pre-existing knowledge, TF alignment methods mitigate this issue. (5) *Stylistic or Normative Alignment*: Empirical evidence from URIAL demonstrates that when the goal is to influence stylistic elements, e.g., tone, politeness, or to ensure ethical compliance without altering factual content, TF alignment can achieve comparable or superior performance of FT alignment purely through ICL with base LLMs.

However, this does not mean that one method can be applied to all the above scenarios, for every method has its own strengths and limitations. Therefore, Table 2 systematically compares the underlying characteristics, compatibility (model accessibility and plug-and-play capability), storage and real-time efficiency, and generalization of different TF alignment methods. We can observe that all in-decoding TF alignment methods are inapplicable to closed-source LLMs, yet they exhibit strong generalization capability. Additionally, TF alignment methods introduce latency due

Category	Characteristic	Compatibility		Efficiency		Generalization
		Model Accessibility	Plug-and-Play	Storage Efficiency	Real-Time Efficiency	
Pre-Decoding TF Alignment	Simple Prompt Engineering	✔	✔	✔	✗	✗
	Enhanced Prompt Strategy	✔	✔	✔✗	✗	✗
	Detector-Based	✔	✔	✔	✗	✔
In-Decoding TF Alignment	Hidden States Adjustment	✗	✗	✔	✗	✔
	Logits Difference Calculation	✗	✗	✔	✗	✔
	Guidance-Based	✗	✗	✔	✗	✔
	Dynamic Search Strategy	✗	✗	✗	✗	✔
Post-Decoding TF Alignment	Filtering/Correcting Outputs	✔	✔	✔	✗	✔

Table 2: A Comparative Analysis of Different TF Alignment Methods. (Note: ✔ and ✗ denote the method has/hasn't the corresponding property. ✔✗ denotes differential storage requirements, where enhanced prompt strategies such as RAG-based architectures or Chain-of-Thought augmented frameworks necessitate storage allocation, whereas other methods remain storage-free.)

to increased prompt length, inference-time adjustments, or input-output detection, resulting in suboptimal real-time efficiency.

We hope these comparisons can enable practitioners to gain a nuanced understanding of each approach and make well-informed decisions when selecting alignment strategies.