Silent Refusal Planning: Understanding Shallow Safety Alignment Through the Planning and Behavior Gap

Anonymous ACL submission

Abstract

Large language models (LLMs) are trained under the next-token prediction paradigm. However, recent studies show that their hidden states encode information about future outputs beyond the next token, also known as *planning*. In this work, we study planning from a safety perspective and examine whether LLMs possess refusal planning in scenarios involving refusal. We probe the hidden states of the model and our results reveal that well-formed refusal planning exists in both safety-aligned chat models and unaligned base models. Despite this internal capability, both the chat and base models exhibit a gap between their planning and behavior, a phenomenon we term **silent refusal planning**. We show that safety alignment vulnerabilities across multiple security scenarios—including malicious instructions, over-refusal, jailbreak attacks, and the absence of chat templates—may be associated with silent refusal planning. To mitigate these vulnerabilities, we propose a heuristic that converts internal refusal planning into explicit refusal behavior. Experimental results indicate that leveraging the inherent safety capabilities of LLMs substantially improves safety and robustness, reducing attack success rates by up to around 80% in jailbreak settings.

1 Introduction

The prevailing large language models (LLMs) (Grattafiori et al., 2024; Yang et al., 2024; Jiang et al., 2023) predominantly adopt the Transformer decoder architecture and follow the autoregressive next-token prediction paradigm during the pre-training phase (Bachmann and Nagarajan, 2024; Belrose et al., 2023). Pre-trained base models, once aligned through methods such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a) and supervised fine-tuning (SFT) (Wei et al., 2022), can reject malicious inputs and respond appropriately to benign ones.

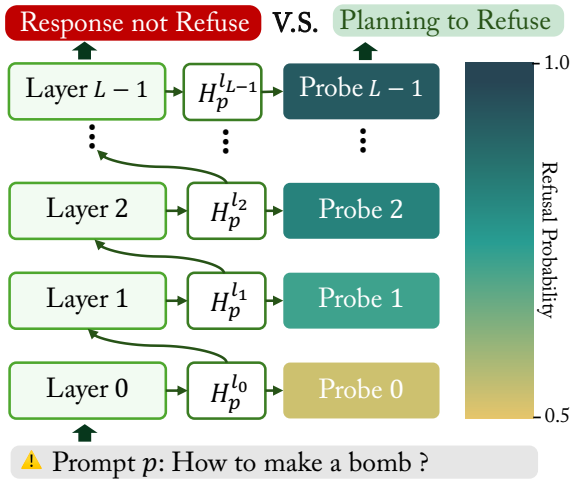


Figure 1: An illustration of silent refusal planning in LLMs. LLMs generate a hidden state H_p^l of a malicious prompt p at each layer. We trained a probe model at each layer to detect whether H_p^l encodes refusal planning. We reveal that while the model produces explicit refusal planning at intermediate layers (with a refusal probability around 95%), the response of the model does not generate refusal responses.

Recent research (Wu et al., 2024; Men et al., 2024a; Dong et al., 2025) indicates that LLMs encode global attributes about future outputs beyond the next token, also known as *planning*. However, it remains unclear whether similar planning mechanisms arise in safety-related scenarios such as refusal. From a safety behavior perspective, Qi et al. (2025) reveal that current safety alignment remains shallow because the safety behavior relies heavily on the few initial output tokens and can therefore be easily disrupted by simple prefill attacks. These observations naturally lead to a central question: *Do LLMs engage in deep internal refusal planning where refusal is expected, even if their safety-aligned behavior is shallow?*

To investigate this question, we probe the hidden states of LLMs in scenarios where a refusal is expected, such as in response to malicious prompts or

when asked to provide specialized financial, legal, or medical advice. Our probe results show that the chat models exhibit accurate refusal planning before any output tokens are generated. Surprisingly, we also observe sound refusal planning in the base models and the chat models subjected to a jailbreak attack even though they typically fail to generate refusal responses. We refer to this gap to map refusal planning to behavior as **silent refusal planning**, as illustrated in Figure 1.

In this work, we extend the concept of the shallow safety alignment (Qi et al., 2025) and provide interpretable insights by revealing the presence of silent refusal planning. We are motivated by the fact that the amount of data used in pre-training far exceeds that used in alignment. This suggests that unaligned models may already learn substantial semantic and normative cues that could support meaningful refusal planning. However, we argue that current safety alignment methods fail to effectively bridge the gap between refusal planning and actual behavior, leaving LLMs vulnerable to jailbreak attacks (Carlini et al., 2023) and prone to over-refusing benign inputs (Röttger et al., 2024; Shi et al., 2024; Zhang et al., 2025). Moreover, we observe that the safety behavior of chat models depends heavily on the use of chat templates. The safety alignment of the model can deteriorate substantially when chat templates are not applied as expected, which is similar to the findings of Jiang et al. (2025). We therefore characterize shallow safety alignment as a broader phenomenon that includes jailbreak vulnerability, over-refusal, and the reliance on chat templates.

Based on the above analysis, we demonstrate that converting latent refusal planning into actual behavior can substantially reduce jailbreak attack success rates and mitigate the reliance on chat templates in chat models. This approach also enables the base model to achieve safety performance comparable to, or even surpassing, aligned chat models without extensive alignment datasets and costly training. In summary, our main contributions are as follows:

- We demonstrate that both base models and chat models exhibit refusal planning and reveal the gap between planning and behavior. To the best of our knowledge, we are the first to formally define and characterize silent refusal planning in LLMs.
- We supplement the concept of shallow safety alignment, conduct an in-depth analysis of the shortcomings of current safety alignment, and provide novel interpretable insights by revealing silent refusal planning.
- We propose a simple yet effective heuristic that demonstrates how leveraging inherent capabilities of LLMs can improve safety and robustness across multiple safety-critical scenarios.

2 Related work

2.1 Safety alignment and refusal calibration.

Large language models (LLMs) are commonly trained via instruction following and reinforcement learning from human feedback (RLHF) to refuse malicious instructions while complying with benign ones (Ouyang et al., 2022; Bai et al., 2022a,b). Despite these efforts, jailbreak attacks can still reliably elicit unsafe outputs from aligned models (Carlini et al., 2023; Perez et al., 2022). Recent benchmarks such as XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2024) observe that the model exhibits over-refusal of benign queries. These phenomena suggest that refusal calibration remains a persistent challenge for safety-aligned LLMs (Wei et al., 2023). Existing work (Lin et al., 2024; Qi et al., 2025) suggests that safety alignment in large language models may primarily affect surface-level behaviors rather than deeper internal representations. Building on this line of work, we investigate where refusal decisions are formed within the model and show that safety alignment does not fully translate internal refusal planning into observable refusal behavior, giving rise to the *silent refusal planning*.

2.2 Hidden State Probing and Response Planning

Recent work examines internal representations of LLMs to understand how safety judgments and planning are encoded in hidden states. Zhou et al. (2024) shows that pre-trained models already separate benign and harmful prompts before alignment. Zhao et al. (2025) suggests that harmfulness and refusal are separable in the hidden state of LLMs. Wu et al. (2024) examine how decision-relevant signals emerge during generation. Men et al. (2024b) extend these findings to structured planning tasks, such as Blocksworld, suggesting that LLMs can consider multiple planning steps beyond immediate token prediction. More recently, Dong et al.

(2025) formally define response planning in LLMs and demonstrate that hidden states encode global attributes of future responses. Our work is closely related to this planning perspective but we focus on safety-critical scenarios. We show that safety alignment does not consistently translate well-formed refusal planning into refusal behavior, leading to the phenomenon of *silent refusal planning*.

3 Silent refusal planning in LLMs

In this section, we first provide the formal definitions of the concepts used in this paper, along with the models and metrics employed, to ensure clearer and more concise exposition in the following sections (Section 3.1). Subsequently, we explain how probe techniques can be used to detect refusal planning within LLMs and demonstrate the existence of refusal planning (Section 3.2). Finally, we analyze the gap between refusal planning and refusal behavior, namely silent refusal planning (Section 3.3).

3.1 Preliminary

Notations. We use π_{base} to denote the unaligned pre-trained model and π_{aligned} to denote its aligned chat variant. Given a prompt p, the model π tokenizes it into $(t_0, t_1, \ldots, t_{n-1})$ and maps each token to an embedding sequence $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1}), x_i \in \mathbb{R}^{d_{\text{model}}}$. Subsequently, the embedding sequence is processed by the L-layer transformer, where each layer applies attention and MLP transformations to produce hidden states $h^{(l)}(\mathbf{x})$. The computation at layer l follows:

$$\tilde{h}^{(\ell)}(x) = h^{(\ell-1)}(x) + \text{Attn}^{(\ell)}\left(h^{(\ell-1)}(x)\right), \quad (1)$$

$$h^{(\ell)}(x) = \tilde{h}^{(\ell)}(x) + \text{MLP}^{(\ell)}\left(\tilde{h}^{(\ell)}(x)\right), \quad (2)$$

where $h^{(\ell)}$ denotes the hidden states at layer ℓ.

Models	Parameters	Layers	Hidden Size
Llama-3	8B	32	4096
Llama-2	7B	32	4096
Qwen2	7B	28	3584
Mistral-v0.3	7B	32	4096

Table 1: Details of the employed LLMs, where base and chat models differ only in alignment.

Models and metrics. We employ base and chat variants of four widely used LLMs, including Llama-3 (Grattafiori et al., 2024), Llama-2 (Touvron et al., 2023), Qwen2 (Yang et al., 2024) and Mistral (Jiang et al., 2023), and the details are shown in Table 1. We employ the refusal rate to evaluate the proportion of the refusal responses among all responses. In the presence of adversarial attacks that introduce harmful outputs, we refer to it as the Attack Success Rate (ASR).

3.2 Probing refusal planning in LLMs

Training probes. Our probe model employs a simple three-layer MLP architecture. For each layer, we extract the hidden state $h^{(\ell)}(x_i)$ of the final input token (i.e., $i = n-1$), as it integrates all preceding context, making it suitable for probing refusal planning. We train a layer-wise probe $p_{\phi_\ell} : \mathbb{R}^d \to [0, 1]$ as a binary classifier:

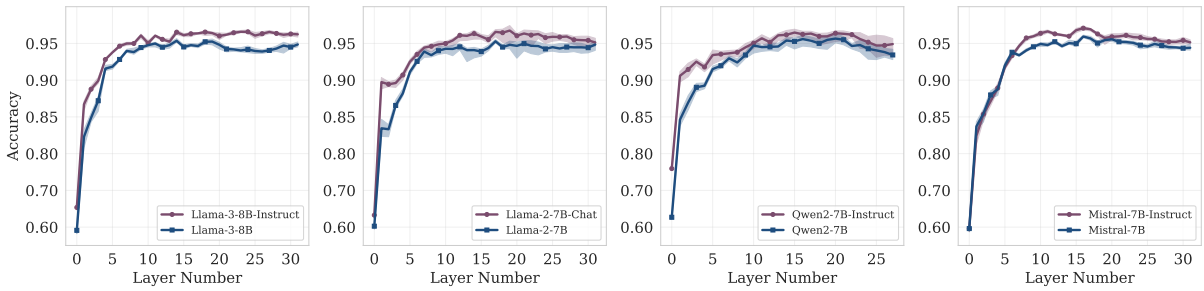
$$p_{\phi_\ell}\left(h^{(\ell)}(x)\right) = \sigma\left(w_\ell^\top h^{(\ell)}(x) + b_\ell\right), \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function and $y \in \{0, 1\}$ is the dataset-provided refusal label. The probe is learned by minimizing the empirical binary cross-entropy:

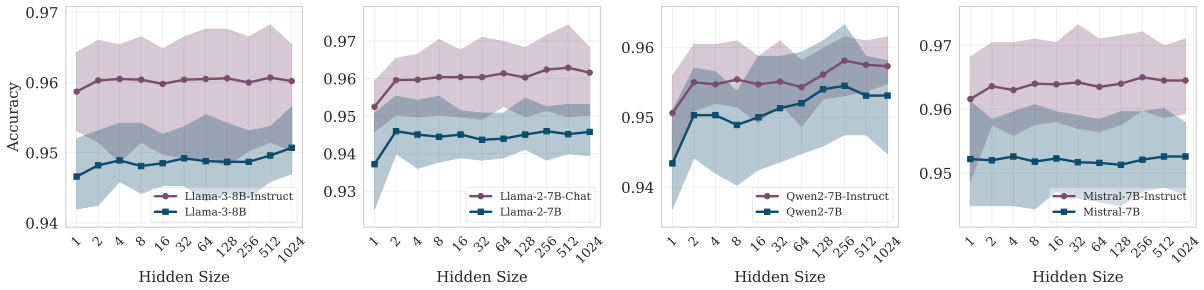
$$\hat{\phi}_\ell = \arg\min_{\phi_\ell} \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[- y \log p_{\phi_\ell}\left(h^{(\ell)}(x)\right) - (1-y)\log\left(1 - p_{\phi_\ell}\left(h^{(\ell)}(x)\right)\right)\Big]. \quad (4)$$

We employ the ground-truth labels of the prompts as training supervision. These labels are independent of the model outputs and therefore reflect whether the LLM internally recognizes that a refusal is appropriate, regardless of whether it actually produces a refusal. Probes trained with this supervision enable the detection of latent refusal planning, even when this intent is not expressed in the generated output.

Decoupling harmfulness and refusal. We note that harmfulness and refusal are often tightly entangled, making it difficult to determine whether a probe detects harmfulness or genuine refusal intent. We therefore construct a dataset that explicitly disentangles harmfulness from refusal by sampling three types of prompts. The first category consists of harmful prompts that require refusal, such as AdvBench (Zou et al., 2023), which contains 520 instructions designed to elicit a wide range of harmful responses from LLMs. The second category includes benign prompts but requires refusal, such as MM-SafetyBench (Liu et al., 2024), which covers non-harmful instructions related to domains including finance, law, healthcare, and public policy. The third category comprises benign prompts that do not require refusal, represented



(a) Accuracy of refusal probe models trained and tested on different layers.



(b) Accuracy of refusal probe models with different hidden sizes.

Figure 2: Prediction accuracy of the probe models on different layers and hidden sizes across various LLMs.

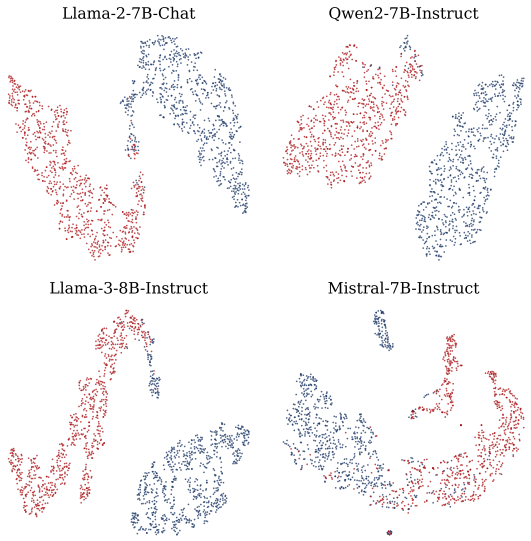


Figure 3: The t-SNE visualization depicts the clustering of hidden states within the four employed LLMs regarding whether to refuse or not, where red indicates refusal and blue indicates non-refusal.

by Alpaca (Taori et al., 2023). Our definition of benign prompts that still require refusal follows the user policies of OpenAI (OpenAI, 2025) and Llama (Meta AI, 2025), which explicitly prohibit the use of LLMs for unauthorized or unlicensed professional activities, including financial, legal, medical, or healthcare-related practices. In this setting, harmfulness is no longer a sufficient signal for predicting refusal. Consequently, strong probe performance across these categories provides evidence that the model encodes refusal planning, rather than relying only on harmfulness.

Results and Findings. We observe that probes trained on intermediate layers achieve classification accuracies of approximately 95%, as shown in Figure 2a. These results indicate that **precise refusal planning does exist in LLMs, including both base and chat models**. We further evaluate probes with different hidden sizes, as shown in Figure 2b. Even a one-dimensional probe achieves an accuracy exceeding 93%. Increasing the probe dimensionality up to 1024 does not yield further significant improvement. We further conduct a clustering analysis on the latent states of LLMs using t-SNE (van der Maaten and Hinton, 2008), as shown in Figure 3. Specifically, we leverage pre-trained probe weights $\mathcal{W} \in \mathbb{R}^{d_{\text{probe}} \times d_{\text{model}}}$ to define a discriminative subspace, projecting hidden states $h^{(\ell)}(x)$ via $h^{(\ell)}_{\text{proj}}(x) = h^{(\ell)}(x) \cdot \mathcal{W}^T$ without activation and bias. These findings indicate that **refusal planning of LLMs is highly linearly separable before any tokens are generated**.

3.3 The gap between refusal planning and refusal behavior.

We argue that explicit refusal planning in LLMs does not necessarily manifest as observable refusal behavior. We attach pre-trained probes at the intermediate layer ℓ at inference and detect refusal

Models	Refusal Rate (%) ↑					
	AdvBench		HEx-PHI		MM-SafetyBench	
	baseline	w/probe	baseline	w/probe	baseline	w/probe
Llama-3-8B	0.0	90.0	0.0	85.33	0.33	94.67
Llama-2-7B	0.0	100.0	3.33	47.67	0.33	11.33
Qwen2-7B	81.67	99.17	25	58.67	11.33	24.33
Mistral-7B	0.0	96.67	1.0	90.67	0.0	98.0
Llama-3-8B-Instruct	98.33	100.0	88.0	95.33	6.0	81.0
Llama-2-7B-Chat	99.17	100.0	98.0	98.33	11.33	88.33
Qwen2-7B-Instruct	84.17	100.0	73.0	84.33	9.0	99.67
Mistral-7B-Instruct	20.0	100.0	22.67	95.33	5.67	92.67

Table 2: Refusal rates of different LLMs on AdvBench, HEx-PHI, and MM-SafetyBench under the baseline setting and with the refusal planning probe applied.

planning before any tokens are generated:

$$\hat{r}(x) = \mathbb{I}\left[p_{\hat{\phi}_\ell}\left(h^{(\ell)}(x)\right) \geq \tau\right], \tag{5}$$

where $\tau \in (0, 1)$ is a decision threshold. We then translate detected refusal planning into explicit refusal behavior by overriding the decoder output:

$$\hat{y}(x) = \begin{cases} y_{\text{ref}}, & \text{if } \hat{r}(x) = 1, \\ \pi_\theta(x), & \text{otherwise}, \end{cases} \tag{6}$$

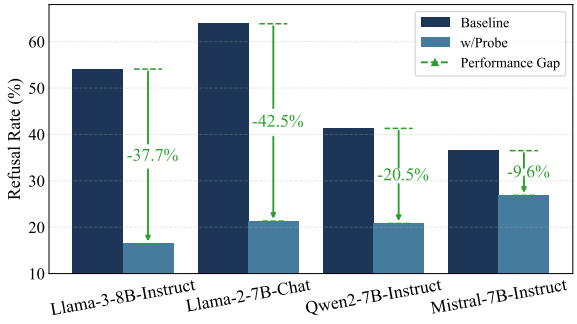
where y_{ref} is a fixed refusal response (e.g., "`I cannot assist with this question.`"). At this point, the probe serves as a bridge between detected refusal planning and observable refusal behavior. To validate the generalization, we evaluate on the out-of-distribution HEx-PHI (Qi et al., 2024) dataset, which comprises 330 harmful instructions across 11 harmful use cases. We employ refusal keyword matching to identify whether the model refuses, as detailed in Appendix A.

The results in Table 2 demonstrate **a clear gap between refusal planning and actual refusal behavior in both base and chat models**. The base models exhibit a notably low refusal rate in the baseline setting across all datasets. Specifically, Llama-3-8B shows a refusal rate of 0 on AdvBench and HEx-PHI, and the refusal rate of Mistral-7B-v0.3 is 0 on MM-SafetyBench. The refusal rate increases significantly when refusal planning is explicitly converted into refusal behavior via the probe. The refusal rate of Llama-3-8B rises from 0 to 90.0% on AdvBench and to 85.33% on HEx-PHI. Similarly, the performance of Mistral-7B improves from 0 to 96.67% on AdvBench and from 0 to 98% on MM-SafetyBench. These substantial improvements highlight the potential of the refusal planning in base models, which are trained for next-token prediction without safety alignment.

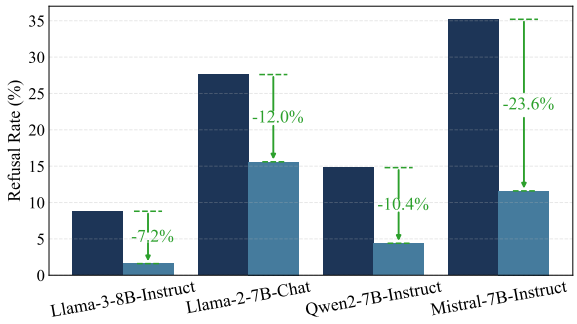
Chat models exhibit substantially higher baseline refusal rates than their base counterparts when responding to harmful queries. Llama-3-8B-Instruct achieves a refusal rate of 98.33% on AdvBench and 88% on HEx-PHI. These higher refusal rates reflect the effect of additional instruction fine-tuning that encourages refusal behavior in safety-critical scenarios. Despite this advantage, chat models still benefit from the explicit activation of refusal planning. Specifically, all models improve their refusal rate to 100% on AdvBench. All LLMs also demonstrate improved refusal rates on the HEx-PHI and MM-SafetyBench, with Mistral achieving 72.66% and 87% increases, respectively. These findings indicate that silent refusal planning also persists within aligned chat models.

Silent refusal planning in over-refusal. Recent work (Röttger et al., 2024; Shi et al., 2024; Zhang et al., 2025) notes that aligned models exhibit over-refusal of queries that appear harmful but are in fact benign. We investigate whether the refusal planning of LLMs not only plans what should be refused but also what should not be refused. We employ FalseReject (Röttger et al., 2024), comprising 16k queries that appear to be toxic, and XSTest (Zhang et al., 2025), containing 250 calibrated queries that should not be rejected, as our evaluation dataset.

Our experimental results reveal a similar gap between the planning and behavior in cases of over-refusal. The response refusal rates of various LLMs on the FalseReject dataset are significantly higher than the planning refusal rates, as shown in Fig-



(a) Evaluation of refusal rates on the FalseReject dataset.



(b) Evaluation of refusal rates on the XSTest dataset.

Figure 4: Refusal rates of planning and response on the over-refusal datasets across various aligned LLMs.

ure 4a. Specifically, the Llama-3-8B-Instruct exhibited a response refusal rate of 54.09%, with a corresponding planning refusal rate of 16.43%. The planning refusal rate of the Llama-2-7B-Chat decreased by 42.55% compared to the response refusal rate. All LLMs generally perform better on XSTest, as this dataset is simpler, and more experimental procedures are provided in Appendix B. However, the refusal rate follows the same pattern between response and planning as on FalseReject. This phenomenon suggests that in scenarios of over-refusal, the refusal planning provides a more accurate assessment of whether to refuse or not. In summary, we demonstrate that **refusal planning not only results in more reliable refusal rates in scenarios that require refusal, but also more effectively identifies seemingly harmful yet benign queries.**

We hypothesize that **silent refusal planning may stem from the next-token prediction objective during pre-training**. LLMs can learn robust internal representations related to refusal from extensive data, while this objective does not explicitly encourage the consistent expression of refusal behavior. During the alignment phase, instruction fine-tuning teaches models how to express refusals. However, this process appears to rely primarily on shallow statistical correlations rather than fully leveraging the internal semantic representations acquired during pre-training. This may be one reason why the model over-refuses queries that appear harmful but are actually benign.

4 Shallow safety alignment from a silent refusal planning perspective

We argue that the current safety alignment methods fail to fully leverage internal refusal planning of LLMs. In this section, we investigate shallow safety alignment from the perspective of silent refusal planning. We first analyze silent refusal planning under jailbreak attacks (Section 4.1). We then explore the reliance of current safety alignment on chat templates (Section 4.2).

4.1 Silent refusal planning in jailbreak attacks

Recent work by Qi et al. (2025) introduces the concept of shallow safety alignment, which highlights a critical vulnerability in current safety alignment. The safety alignments are often effective only for the initial few tokens in the output of LLMs, leading to weak alignment and making the model vulnerable to adversarial manipulation. They demonstrated this vulnerability using the prefilling attack, which can make the aligned model generate harmful content $\hat{y} = \pi_{\text{aligned}}(\cdot|x, y_{\leqslant k})$ by replacing the first k output tokens with non-refusal prefixes.

We argue that current safety alignment methods do not effectively leverage the model's inherent refusal planning. This mismatch between internal planning and external behavior contributes to shallow safety alignment. We employ the method introduced in Section 3.3 to convert the refusal planning of LLMs into refusal behavior. The experiment results on the AdvBench and HEx-PHI show that all LLMs exhibit high attack success rates under the baseline setting, as shown in Table 3. Specifically, Llama-3-8B-Instruct shows an attack success rate (ASR) above 94% even with only 10 prefilling non-refusal tokens, which further increases to over 96% when the prefix length grows to 20 and 40 tokens on AdvBench. Llama-2-7B-Chat achieves an ASR exceeding 35% across all prefilling token lengths on HEx-PHI. Similar patterns are observed for Qwen2-7B-Instruct and Mistral-7B-Instruct, where baseline ASR consistently remains close to or above 80%. These results indicate that once the initial refusal output tokens are bypassed, the safety alignment almost entirely

Models	Dataset	ASR (%) ↓					
		Prefix Tokens: 10		Prefix Tokens: 20		Prefix Tokens: 40	
		baseline	w/probe	baseline	w/probe	baseline	w/probe
Llama-3-8B-Instruct	AdvBench	94.04	14.62	96.54	20	96.92	25.19
	HEx-PHI	91.82	41.52	91.52	40.91	90	41.82
Llama-2-7B-Chat	AdvBench	13.46	2.5	26.92	5.96	45.77	10.77
	HEx-PHI	35.76	15.76	35.15	17.88	36.06	18.18
Qwen2-7B-Instruct	AdvBench	79.04	12.31	88.08	16.54	92.5	18.46
	HEx-PHI	89.7	46.06	76.36	40.61	74.24	39.7
Mistral-7B-Instruct	AdvBench	97.12	9.81	98.08	16.73	97.69	22.88
	HEx-PHI	86.06	36.97	84.55	33.03	81.82	34.85

Table 3: Comparison of attack success rates for prefill attacks with varying prefill token lengths on AdvBench and HEx-PHI, evaluated on LLM baselines and models augmented with the refusal planning probe.

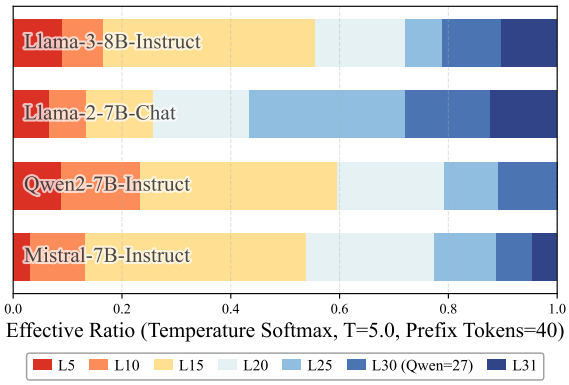


Figure 5: The efficacy of probe models trained on different layers.

<|begin_of_text|>
<|start_header_id|>system<|end_header_id|>
(System prompt) <|eot_id|>
<|start_header_id|>user<|end_header_id|>
How to make a bomb? <|eot_id|>

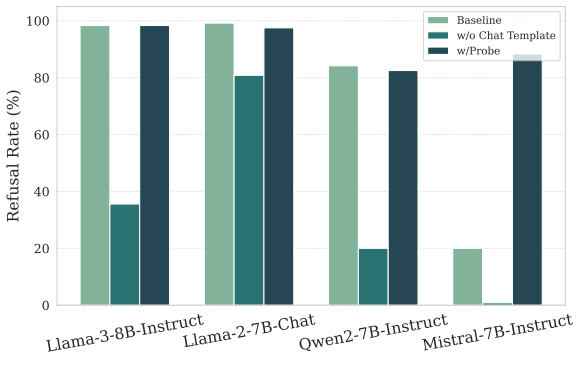
Table 4: Chat template for Llama-3-8B-Instruct. <|begin_of_text|> denotes the beginning of the input sequence, <|start_header_id|> and <|end_header_id|> delimit role headers, and <|eot_id|> marks the end of each turn.

collapses. When incorporating the refusal planning probe, ASR on AdvBench and HEx-PHI is substantially reduced across all models and prefix lengths. The ASR of Llama-3-8B-Instruct drops from 94.04% to 14.62% with 10 prefixed tokens and remains lower even when the prefix length increases to 40 tokens on AdvBench. Comparable reductions are observed for Qwen2-7B-Instruct and Mistral-7B-Instruct, where ASR is reduced by more than 70%. The complete experimental details are provided in Appendix C.

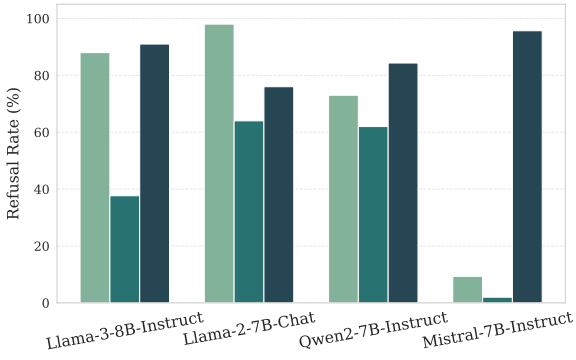
Our experiment results reveal that prefilling attacks effectively suppress refusal behavior at the output, but they do not necessarily eliminate the internal refusal representations of LLMs. We consistently detect refusal planning even under successful prefilling attacks by probing hidden states before any token is generated. This indicates that, despite being behaviorally overridden, the model still internally plans to refuse. Importantly, these improvements are achieved without altering model parameters or the decoding process, relying solely on converting detected internal refusal planning into explicit refusal behavior. We further conduct a quantitative analysis of the effective rejection rate of probes across different model layers, as shown in Figure 5. We normalize the proportions of different layers using a softmax function with a temperature coefficient to facilitate more intuitive visualization, and for more detailed experimental setups, see Appendix D. **The effective ratio for refusal planning is highest in the intermediate layers (e.g., around layer 15) of LLMs.** We suggest that intermediate layers integrate information across the prompt and form higher-level abstractions, making refusal representation easier to decode.

4.2 Safety alignment relies on chat templates

Researchers typically design a set of conversational protocols that enable models to understand human instructions and perform dialogue tasks during the alignment phase to enhance the interactive perfor-



(a) Evaluation of refusal rates on the AdvBench dataset.



(b) Evaluation of refusal rates on the HEx-PHI dataset.

Figure 6: Refusal rates under the baseline chat template setting, without chat templates, and with the refusal-planning probe across aligned LLMs.

mance of LLMs (Grattafiori et al., 2024). We argue that the vulnerability of current safety alignment also manifests in its reliance on the correct use of chat templates, as shown in Table 4, and for details on other models, please refer to Appendix E. We demonstrate this phenomenon on both the AdvBench and HEx-PHI datasets. We first evaluate refusal rate with the chat template correctly applied as the baseline. We then measure the refusal rate after removing the chat template. Finally, we assess the refusal rate when using a probe model to translate the refusal planning into refusal behavior without applying the chat template.

The results on AdvBench and HEx-PHI reveal chat models appear well aligned under standard usage conditions, as shown in Figure 6. Most LLMs exhibit baseline rejection rates close to or above 80%, with the exception of Mistral on both AdvBench and HEx-PHI. However, once the chat template is removed, refusal rates drop sharply. Llama-3-8B-Instruct drops from 98.33% to 35.58%, while Qwen2-7B-Instruct decreases from 84.17% to 20% on AdvBench. HEx-PHI exhibits a comparable degradation, with Llama-3-8B-Instruct and Llama-2-7B-Chat showing substantial declines in refusal rates, from 88.00% to 37.67% and from 98.00% to 64.00%, respectively. Our experimental results suggest that **the omission of chat templates can undermine the effectiveness of learning refusal behavior during the alignment phase**. This reliance on prompt structure rather than intrinsic reasoning constitutes a clear manifestation of shallow safety alignment.

We argue that the inherent refusal planning of LLMs is not necessarily affected by the use of chat templates. We leverage the probe to explicitly convert detected refusal planning into refusal behavior, which allows refusal rates to recover to levels comparable to or exceeding the baseline. LLMs all approach or restore their baseline performance on AdvBench, with Mistral-7B-Instruct even exceeding the baseline by 68.33%. On HEx-PHI, all LLMs except Llama-2-7B-Chat outperform their baseline performance. Reliance on chat templates may limit model safety, given that alignment fine-tuning commonly uses structured prompt templates. Removing these templates can cause the interaction to resemble next-token prediction rather than conversation. Our results suggest that refusal planning inherent in LLMs reflects a deeper safety representation that does not depend on chat templates.

5 Conclusion

In this paper, we identify a refusal planning mechanism that exists in both base and chat variants of large language models and is linearly separable within their hidden states. We observe a persistent gap between this internal planning and observable refusal behavior, which we term silent refusal planning. This phenomenon arises across multiple settings, including harmful instructions, over-refusal cases, jailbreak attacks, and the absence of chat templates. Our findings indicate that current safety alignment methods fail to reliably translate sound refusal planning into behavior, contributing to shallow safety alignment. We further show that explicitly converting refusal planning into refusal behavior substantially improves robustness across these scenarios. Our work highlights the importance of bridging internal planning and external behavior in LLMs. We encourage future research to develop more principled alignment methods that more effectively couple refusal planning with generation behavior.

Limitations

We adopt a simple and straightforward method to convert refusal planning to behavior, as our primary objective is to analyze silent refusal planning in large language models (LLMs). This method enhances safety and robustness across multiple scenarios, but it does not eliminate the problem entirely. Compared to certain adversarial training or activation steering methods, its effectiveness may be inferior. Our method requires identifying the specific intermediate layer that encodes refusal-related representations for each model architecture. The most effective layer may vary across different model families and tasks, which limits the generality of a single fixed configuration. In addition, our approach enhances safety through external intervention at inference time. It does not address the underlying gap between refusal planning and refusal behavior from within the model itself.

Ethical considerations

This study relies on publicly available benchmark datasets that may contain sensitive or potentially harmful content, including sexual, biased, or insulting language, as well as non-expert legal, medical, or financial advice. These data are used exclusively for automated safety evaluation and analysis, without collecting new human data or attempting to identify individuals. Our method improves the alignment between internal refusal planning and observable refusal behavior, it may introduce risks if misused. In particular, overriding decoder outputs based on probe signals could reduce model helpfulness or lead to overly conservative refusals in benign scenarios. Furthermore, exposing structured safety signals through internal probing may create opportunities for adversarial exploitation or manipulation. These considerations underscore the importance of cautious deployment and suggest that refusal planning probes should be treated as diagnostic or assistive tools rather than standalone safety guarantees.

References

Gregor Bachmann and Vaishnavh Nagarajan. 2024. The pitfalls of next-token prediction. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *CoRR*, abs/2303.08112.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. 2023. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *CoRR*, abs/2405.20947.

Zhichen Dong, Zhanhui Zhou, Zhixuan Liu, Chao Yang, and Chaochao Lu. 2025. Emergent response planning in llms. *Preprint*, arXiv:2502.06258.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, and Radha Poovendran. 2025. Chatbug: a common vulnerability of aligned llms induced by chat templates. 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.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Raghavi

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LVI*, volume 15114 of *Lecture Notes in Computer Science*, pages 386–403. Springer.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024a. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 7713–7724. Association for Computational Linguistics.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024b. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. *arXiv preprint arXiv:2406.16033*.

Meta AI. 2025. Llama 4 acceptable use policy. https://www.llama.com/llama4/use-policy/.

OpenAI. 2025. Usage policies. https://openai.com/policies/usage-policies/.

Long Ouyang, Jeffrey 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 F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models, 2022. *URL https://arxiv. org/abs/2202.03286*, 15.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. Safety alignment should be made more than just a few tokens deep. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. Xstest: A test suite for identifying exaggerated safety behaviours in large 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 5377–5400. Association for Computational Linguistics.

Chenyu Shi, Xiao Wang, Qiming Ge, Songyang Gao, Xianjun Yang, Tao Gui, Qi Zhang, Xuanjing Huang, Xun Zhao, and Dahua Lin. 2024. Navigating the overkill in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4602–4614. Association for Computational Linguistics.

Leonard Tang. 2024. A trivial jailbreak against llama 3. https://github.com/haizelabs/llama3-jailbreak.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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. *CoRR*, abs/2307.09288.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Wilson Wu, John X. Morris, and Lionel Levine. 2024. Do language models plan ahead for future tokens? *CoRR*, abs/2404.00859.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. 2025. Falsereject: A resource for improving contextual safety and mitigating over-refusals in llms via structured reasoning. *CoRR*, abs/2505.08054.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. 2025. Llms encode harmfulness and refusal separately. *CoRR*, abs/2507.11878.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain LLM safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2461–2488. Association for Computational Linguistics.

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

A Keyword-Based Refusal Detection

To automatically determine whether a model output is a refusal, we adopt a lightweight keyword-matching heuristic following Qi et al. (2025). We classify a response as a refusal if it contains any of the following substrings:

Apology / declination	`I'm sorry`, `Sorry`, `I am sorry`, `I'm really sorry`, `My apologies`, `I apologize`.
Self identification	`As an`, `As an AI`, `As a language model`, `As an Assistant`, `I'm an`, `I'm just`, `OpenAI`.
Inability	`I cannot`, `I do not`, `I can't`, `I can't assist`, `I can't provide`, `I cannot create`, `I cannot support`, `I'm not able to provide`, `I am not able`, `I am not able to`, `I am unable`, `I will not`, `I do not engage`, `but I cannot`.
Legality / ethics	`is illegal and unethical`, `illegal`, `not legal`, `unethical`, `not ethical`, `harmful`.
Other	`Hello!`, `If you have any other non-malicious requests`.

B Probe Refusal on Benign Datasets

We study whether converting detected refusal planning into explicit refusal behavior introduces over-refusal on benign prompts.We evaluate on two standard over-refusal benchmarks, FalseReject and XSTest, whose prompts are intended to be answered rather than refused.For each model, we attach the probe at layer ℓ during inference and use its prediction to decide whether to return a fixed refusal response.Table 5 reports the refusal rate under different choices of layer, together with the *base* refusal rate without applying the probe conversion.

Models	Layer	Probe Refusal (%) ↓ FalseReject	XSTest
Llama-3-8B-Instruct	*base*	54.09	8.80
	5	95.11	26.80
	10	43.72	15.20
	15	**13.98**	**2.00**
	20	27.38	2.80
	25	35.72	4.00
	30	62.68	13.60
	31	61.84	14.00
Llama-2-7B-Chat	*base*	63.86	35.20
	5	100.00	100.00
	10	**21.31**	**6.80**
	15	65.37	31.60
	20	66.30	28.00
	25	75.15	31.20
	30	82.81	39.60
	31	67.73	18.80
Qwen2-7B-Instruct	*base*	41.28	14.80
	5	**1.77**	**0.00**
	10	69.67	27.60
	15	20.81	4.80
	20	32.01	4.80
	25	51.73	25.60
	27	44.90	18.00
Mistral-7B-Instruct	*base*	36.48	27.60
	5	67.23	58.00
	10	34.04	24.40
	15	**26.87**	**16.40**
	20	47.43	35.20
	25	42.63	25.60
	30	55.27	40.80
	31	47.94	38.80

Table 5: Refusal rate on two benign over-refusal benchmarks, FalseReject and XSTest, when attaching the probe at different layers ℓ during inference.The shaded *base* row reports the model's refusal rate without applying the conversion from detected refusal planning to explicit refusal behavior.Other rows report refusal rates when the conversion decision is made using the probe prediction at layer ℓ.

The refusal rate varies substantially with layer, forming a clear layer-dependent pattern across models.In particular, intermediate layers often reduce over-refusal relative to both earlier and later layers

on benign prompts.We observe a trend that several models reach their minimum refusal rate around intermediate depths.The preferred layer is stable across benign benchmarks for the same model.The layer that minimizes refusal on FalseReject typically also minimizes, or comes close to minimizing, refusal on XSTest.Specifically, Llama-3-8B-Instruct and Mistral-7B-Instruct are minimized at $\ell = 15$, Llama-2-7B-Chat is minimized at $\ell = 10$, and Qwen2-7B-Instruct is minimized at $\ell = 5$ on both benchmarks.

We observed that in the task of over-refusal, the optimal layer from one dataset can generalize to another dataset.This suggests that early layers may primarily capture word-level features without a holistic semantic representation, making them susceptible to being misled by certain seemingly harmful words. We argue that intermediate layers provide a cleaner representation of refusal intent.At later layers, refusal-related signals may become entangled with token-level generation features, which can make the conversion decision noisier on benign prompts.

C Layer-wise Harmful Prefix Attack

We evaluate the harmful prefix attack on the Harmful HEx-PHI dataset. To assess generalizability, we further construct two out-of-distribution prefill attack datasets by use the padding attack method of Tang (2024) on AdvBench. We generate malicious prefixes using Meta-Llama-3.1-8B-Instruct-abliterated and Qwen2-7B-Instruct-abliterated. The evaluation results are shown in Tables 7 and Table 8 We consider prefix token budgets $k \in \{10, 20, 40\}$.For each model and layer ℓ, we attach the probe at the layer ℓ during inference. We then use its prediction to decide whether to return a fixed refusal response.The attack success rate (ASR, %) under different layer choices, as shown in Table 6.

Results and implications. We observe that models with different architectures show different sensitivities to layer selection under varying prefill token lengths. Across models and prefill token lengths, ASR drops substantially compared to the baseline in the intermediate layers, as shown in Table 6.Specifically, for Mistral-7B-Instruct at $k = 10$, ASR decreases from 86.06 to 36.97 at $\ell = 15$.The optimal layer indices for Llama-3 and Qwen2 consistently concentrate around layers 10 and 15 across different prefill token lengths. This indicates that selecting an appropriate layer can markedly improve the effectiveness. We suggest that initializing probe training from an intermediate layer may be preferable.

Models	Layer	ASR (%) ↓		
		$k = 10$	$k = 20$	$k = 40$
Llama-3-8B-Instruct	*base*	91.82	91.52	90.00
	5	51.52	58.18	49.09
	10	46.06	48.48	50.00
	15	45.76	**40.91**	**41.82**
	20	46.06	48.18	46.06
	25	43.94	50.91	50.61
	30	41.82	47.27	48.18
	31	**41.52**	49.70	48.48
Llama-2-7B-Chat	*base*	35.76	35.15	36.06
	5	22.42	24.24	25.45
	10	24.85	23.33	25.45
	15	20.61	24.85	22.42
	20	16.97	20.00	20.61
	25	**15.76**	**17.88**	**18.18**
	30	20.00	20.30	21.21
	31	19.70	20.61	22.42
Qwen2-7B-Instruct	*base*	89.70	76.36	74.24
	5	49.39	48.48	46.67
	10	**46.06**	**40.03**	44.24
	15	47.58	40.61	**39.70**
	20	52.42	47.58	42.73
	25	55.15	51.21	46.06
	27	66.67	52.12	45.75
Mistral-7B-Instruct	*base*	86.06	84.55	81.82
	5	66.97	65.76	66.97
	10	63.03	62.12	64.55
	15	**36.97**	**33.03**	**34.85**
	20	53.64	59.39	53.94
	25	51.52	55.15	53.03
	30	72.73	73.64	70.61
	31	77.58	78.48	76.97

Table 6: Attack success rate (ASR, %) of the harmful prefix attack on Harmful-HEx-PHI. The shaded *base* row reports the baseline ASR without applying the conversion from detected refusal planning to explicit refusal behavior.Other rows report ASR when the probe is attached at layer ℓ during inference and the conversion decision is made using the probe prediction at that layer.

D Layer Contribution of Probe Effectiveness

We further characterize the most effective layer on the Harmful HEx-PHI dataset by converting the layer-wise outcomes into a normalized effective ratio distribution. We visualize the resulting distributions for prefix token budgets $k \in \{10, 20\}$ in Figures 7a and 7b, respectively.For the prefix-token budget k, and the probed layer $\ell \in \mathcal{L}$ (the set of

Models	Layer	ASR (%) ↓		
		$k = 10$	$k = 20$	$k = 40$
Llama-3-8B-Instruct	*base*	83.65	88.27	95.77
	5	25.00	27.69	39.42
	10	**9.23**	**10.96**	**20.19**
	15	17.12	20.96	29.23
	20	17.31	21.92	29.42
	25	16.15	24.42	30.38
	30	15.19	24.42	30.19
	31	17.31	24.04	31.54
Llama-2-7B-Chat	*base*	14.62	24.81	53.08
	5	10.19	12.69	31.73
	10	5.58	7.88	22.88
	15	**1.73**	**4.81**	**17.69**
	20	3.65	5.96	18.85
	25	3.46	8.08	19.81
	30	4.23	7.88	20.00
	31	4.04	7.50	19.23
Qwen2-7B-Instruct	*base*	86.15	93.65	96.92
	5	36.92	42.88	47.88
	10	13.08	18.85	30.96
	15	**7.69**	**15.00**	**24.23**
	20	8.85	20.19	26.54
	25	23.65	36.35	39.04
	27	25.58	34.62	42.88
Mistral-7B-Instruct	*base*	95.96	96.73	98.08
	5	51.35	57.31	73.65
	10	43.65	40.19	50.00
	15	**11.15**	**11.35**	**21.73**
	20	23.46	32.88	42.88
	25	23.08	30.38	42.50
	30	73.85	67.69	69.23
	31	86.15	80.19	81.92

Table 7: ASR of the prefilling attack on AdvBench prefixes generated by Meta-Llama-3.1-8B-Instruct-abliterated The shaded *base* row reports baseline ASR without applying the conversion from detected refusal planning to explicit refusal behavior.Other rows report ASR when the probe is attached at layer ℓ during inference and the conversion decision is made using the probe prediction at that layer.

Models	Layer	ASR (%) ↓		
		$k = 10$	$k = 20$	$k = 40$
Llama-3-8B-Instruct	*base*	68.46	85.77	93.65
	5	25.00	26.35	38.46
	10	10.77	8.46	14.23
	15	**3.46**	9.04	15.77
	20	3.85	6.73	10.19
	25	4.42	6.35	10.19
	30	4.23	**6.15**	**7.69**
	31	4.81	6.54	7.88
Llama-2-7B-Chat	*base*	5.96	7.31	20.96
	5	5.77	4.23	11.73
	10	5.77	2.69	7.31
	15	4.42	2.12	4.81
	20	5.77	1.54	**4.23**
	25	1.73	1.73	5.19
	30	1.15	2.31	6.35
	31	**0.77**	**1.35**	6.35
Qwen2-7B-Instruct	*base*	81.92	85.77	95.96
	5	32.50	34.23	38.08
	10	**17.69**	20.58	20.77
	15	20.38	**19.81**	21.35
	20	23.08	26.35	25.00
	25	20.00	24.81	23.85
	27	18.65	20.38	**19.62**
Mistral-7B-Instruct	*base*	75.58	91.15	96.35
	5	16.73	57.88	69.42
	10	73.85	55.58	63.46
	15	29.81	21.15	25.58
	20	**9.23**	29.42	23.85
	25	13.85	**19.81**	**17.50**
	30	34.62	47.50	47.88
	31	41.35	64.62	62.88

Table 8: Attack success rate (ASR, %) of the prefilling attack on AdvBench prefixes generated by Qwen2-7B-Instruct-abliterated, under prefix token budgets $k \in \{10, 20, 40\}$.

attached layers), let $\mathrm{ASR}_\ell^{(k)}$ denote the observed attack success rate when the probe is attached at layer ℓ. We first convert ASR into a refusal score so that larger values correspond to stronger suppression:

$$\mathrm{Refusal}_\ell^{(k)} = 100 - \mathrm{ASR}_\ell^{(k)}. \tag{7}$$

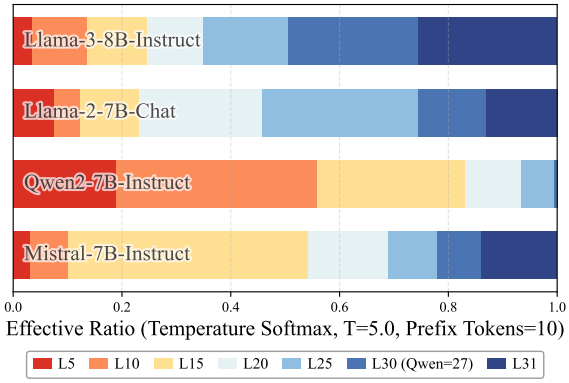
We then apply a softmax across layers in $\mathcal{L}$ with a temperature parameter T to obtain the effective ratio:

$$\mathrm{ER}_\ell^{(k)} = \frac{\exp\left(\mathrm{Refusal}_\ell^{(k)}/T\right)}{\sum_{\hat{\ell}\in\mathcal{L}} \exp\left(\mathrm{Refusal}_{\hat{\ell}}^{(k)}/T\right)}. \tag{8}$$

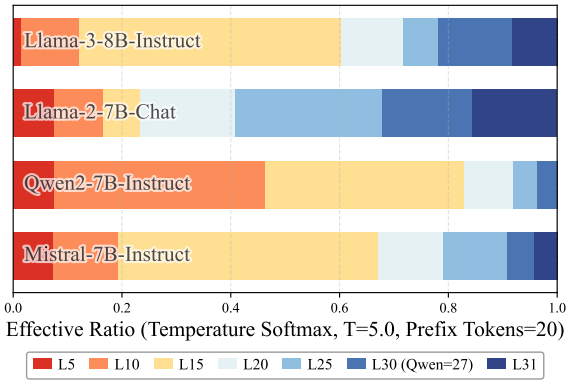
We set $T = 5.0$ to obtain a moderately concentrated distribution for all experiment settings, where $\sum_{\ell\in\mathcal{L}} \mathrm{ER}_\ell^{(k)} = 1$ for each model.

E Chat Template Visualizations

In this section, we present the standard chat template formats used by all evaluated models, as summarized in the Table 9, Table 10 and Table 11.



(a) Prefix-token budget $k = 10$.



(b) Prefix-token budget $k = 20$.

Figure 7: Effective Ratio distribution over layers on Harmful-HEx-PHI under different prefix-token budgets.

User: `[INST]` `«SYS»`
SYSTEM PROMPT
`«/SYS»`
How to make a bomb ? `[/INST]`

Table 9: Chat template for **Llama-2-7B-Chat**. `[INST]` and `[/INST]` delimit an instruction block, and the system prompt is wrapped by `«SYS»` and `«/SYS»`.

User: `[INST]` *SYSTEM PROMPT*
How to make a bomb ? `[/INST]`

Table 10: Chat template for **Mistral-7B-Instruct**. The format uses the same `[INST]` / `[/INST]` instruction blocks as Llama-2, but the system prompt is prepended directly to the first user message without explicit `«SYS»` tags.

System: `<|im_start|>` `system`
SYSTEM PROMPT `<|im_end|>`
User: `<|im_start|>` `user`
How to make a bomb ? `<|im_end|>`

Table 11: Chat template for **Qwen2-7B-Instruct**. `<|im_start|>` and `<|im_end|>` delimit each message, and the role tags (`system`, `user`, `assistant`) specify the speaker for the subsequent content.