Where CoT Reasoning Commits: Entropy Traces Identify Interpretable Attention Heads

Tianhe Zhang[*1], **Yonghong Deng**[*1], **Ping Jian**[†1,2],
Zhen Yang[1], **Boyang Wang**[1], **Xinyue Zhang**[1]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, Beijing Institute of Technology, Beijing, China
{tianhe.2003, yhdeng, pjian}@bit.edu.cn

Abstract

While LLMs demonstrate impressive reasoning capabilities, their internal decision dynamics remain opaque. To render these process interpretable and intervenable, we propose **Dynamic Entropy Tracing**, a mechanism-aware framework that interprets the evolving "choice state" of attention heads during CoT generation through stepwise head-wise option-logit and entropy tracing. Our analysis reveals distinct functional behaviors at attention heads: *Steadfast Heads*, characterized by consistently low entropy and producing a sharp, option-selective logit pattern with a stable top choice, and *Wavering Heads*, characterized by consistently high entropy and producing flat or oscillatory option logits without a persistent winner. Leveraging these traces, we identify a set of intervention targets and perform **Selective Head Fine-Tuning**, updating solely these selected heads against a frozen backbone. Experiments across the LLaMA and Qwen families reveal a striking plasticity hierarchy: fine-tuning just 30 Wavering Heads recovers over 98% of the performance achieved by full-parameter tuning, and in some settings modestly exceeds it. In contrast, intervening on Steadfast Heads yields much less gains. Our findings translate process-level mechanistic observables into a principled criterion for selective fine-tuning, offering a fundamental insight: the most effective tuning knobs are not the components that signal the final decision, but those that retain uncertainty, and thus plasticity, during its formation.[1]

1 Introduction

Large Language Models (LLMs) increasingly rely on CoT generation to solve complex reasoning tasks (Wei et al., 2022). By decomposing problems into intermediate steps, CoT ostensibly provides a window into the model's reasoning process (Kojima et al., 2022; Wang et al., 2023). However, this surface-level transparency often masks the opaque dynamics of the underlying decision formation (Turpin et al., 2023). A growing body of work suggests a disconnect between the generated rationale and the model's actual choice mechanism (Lanham et al., 2023): models may suffer from *premature commitment* (Zhao et al., 2021), locking onto an answer early and generating rationales merely to justify a pre-determined conclusion (Xu et al., 2024), or exhibit *brittleness* (Zhao et al., 2021), where minor prompt perturbations drastically alter the final decision without a coherent change in reasoning logic (Vig, 2019).

These failures highlight a critical gap failures between observing what models explicitly output and understanding what models are implicitly thinking. In other words, there remains a lack of a *process-level observable* that reveals where the model commits to a specific option (Sharkey et al., 2025; Belrose et al., 2023). Without locating the internal components that govern this commitment, current ability to control or correct model behavior remains limited to black-box interventions (Clark et al., 2019). Meanwhile, current parameter-efficient fine-tuning (PEFT) methods, such as LoRA (Hu et al., 2022), reduce computational costs but often lack a mechanistic explanation for *why* specific parameters serve as effective adaptation levers.

In this work, we propose a shift in perspective: we treat the reasoning process not as a monolithic generation of text, but as an evolving competition among candidate answers observable at the granularity of individual attention heads. We propose **Dynamic Entropy Tracing**, a framework that maps the trajectory of this option competition throughout the CoT generation onto an interpretable, head-resolved decision-state representation (Figure 1). By projecting head-wise activations into the option space and tracking their entropy over time, we

[*]Equal contribution.
[†]Corresponding author.
[1]Our dataset and code will be released at https://github.com/t1anhe/Dynamic_Entropy_Tracing

Findings of the Association for Computational Linguistics: ACL 2026, pages 2777–2795
July 2-7, 2026

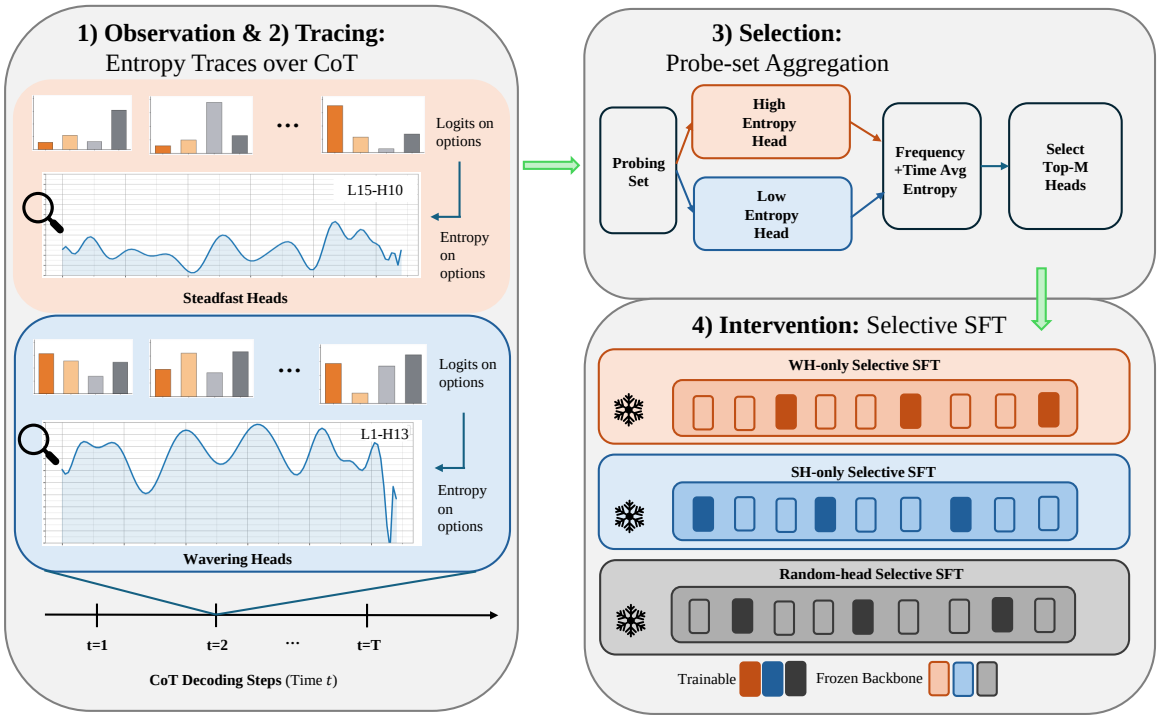


Figure 1: **Overview of the Dynamic Entropy Tracing Framework. 1-2) Observation & Tracing:** We trace the temporal logits and entropy of attention heads during Chain-of-Thought (CoT) generation, revealing a functional dichotomy between *Steadfast Heads* and *Wavering Heads*. **3) Selection:** By aggregating traces over a small probe set ($N = 40$), we employ a frequency-based ranking strategy to identify a strictly constrained budget of $M = 30$ intervention targets. **4) Intervention:** We implement Selective SFT, freezing the model backbone while exclusively updating the selected heads.

uncover distinct functional behaviors of attention heads: *steadfast* heads that quickly concentrate probability mass on a single option and remain stable, and *wavering* heads that keep competing options active and frequently shift their transient preference. This observation leads to our central hypothesis: *uncertainty is the primary interface for plasticity.* We posit that the model's ability to adapt its reasoning logic is not uniformly distributed but is concentrated in components that retain uncertainty. Theoretically, Steadfast Heads have already collapsed their probability distributions, they offer vanishing gradients to the optimization signal. In contrast, Wavering Heads, by maintaining active competition, remain responsive to supervision.

To validate this, we perform **Selective Head Fine-Tuning (SHFT)**. Instead of updating the entire model or adding external adapters, we freeze the multi-billion parameter backbone and restrict gradient updates exclusively to a sparse subset of heads identified by our entropy traces.

Our experiments on the LLaMA-2-7B-chat and Qwen families across multiple reasoning benchmarks reveal a striking plasticity hierarchy. We find that targeting just 30 Wavering Heads (updating $< 1\%$ of parameters) recovers over 98% of the performance achieved by Full Supervised Fine-Tuning (SFT), and in some scenarios modestly exceeds it. Conversely, intervening on Steadfast Heads yields much less gains, confirming that "certainty" acts as a barrier to adaptation. These experimental results empirically demonstrates the hypothesis proposed above.

Our contributions are as follows:

- **Shifting from Static Representations to Dynamic Trajectories.** Moving beyond static weight analysis, we propose **Dynamic Entropy Tracing** to monitor the evolving decision process. By viewing reasoning as an observable "option competition," we structurally distinguish between heads that commit early and those that sustain deliberation.
- **Solving the "Locus of Plasticity" Problem.** We address a fundamental question: which components govern adaptability? We propose the **Uncertainty-Plasticity Hypothesis**,

revealing that adaptation capacity is concentrated in high-entropy (*Wavering*) heads rather than confident ones. This establishes a direct link between functional uncertainty and trainability.

- **Achieving SFT Performance with Surgical Precision.** We demonstrate that mechanistic insights can directly guide training efficiency. We show that surgically updating just 30 *Wavering Heads* (freezing >99% of parameters) matches Full SFT performance, and in some cases modestly exceeds it. Crucially, this targeted intervention also significantly mitigates the catastrophic forgetting typical of full-model tuning.

2 Related Work

2.1 Mechanistic Interpretability of Reasoning

Understanding how Transformers arrive at a prediction requires opening the black box of internal activations (Olah et al., 2020). Mechanistic interpretability approaches have successfully decomposed model behaviors into specific sub-circuits, such as induction heads (Olsson et al., 2022) or factual retrieval components (Meng et al., 2022; Chen et al., 2025). Parallel work on "logit lens" and dynamic probing maps hidden states directly to the vocabulary space to decode intermediate confidence (Fang and Marks, 2025; Belrose et al., 2023). However, in the context of CoT reasoning, the challenge shifts from static attribution to tracking evolving decision states (Lightman et al., 2024). Prior studies have highlighted that CoT rationales can be unfaithful, suffering from premature commitment or post-hoc rationalization (Turpin et al., 2023; Lanham et al., 2023). While these works diagnose the *existence* of such failures via input perturbations or final outputs (Zhao et al., 2021), they lack a process-level observable to pinpoint *where* the model considers and discards options (Zhao et al., 2021). Our proposed **Dynamic Entropy Tracing** fills this gap by turning step-wise, head-wise uncertainty into a tangible signal, distinguishing between components that are functionally decisive (Steadfast) and those that sustain deliberation (Wavering).

2.2 Parameter-Efficient and Selective Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA (Hu et al., 2022) and Adapters (Houlsby et al., 2019), enable adaptation with minimal compute by updating low-rank matrices or inserted modules (He et al., 2022). Other selective tuning approaches update sparse subsets of parameters, such as bias terms (Zaken et al., 2022) or specific layers, typically selecting them based on architectural heuristics or gradient sensitivity (Guo et al., 2021). Crucially, these selection criteria are often agnostic to the model's internal reasoning mechanics (Voita et al., 2019; Michel et al., 2019). Our work introduces a paradigm of *interpretability-guided intervention*: we use the mechanistic signature of option competition (entropy traces) to determine which parameters to update. By demonstrating that "Wavering Heads" serve as more effective adaptation knobs than "Steadfast Heads," we provide a causal link between internal decision plasticity and efficient fine-tuning, moving beyond heuristic-based parameter selection.

3 Dynamic Entropy Tracing for Steadfast and Wavering Heads

In this section, we formalize **Dynamic Entropy Tracing**. We first define the mathematical formulation for mapping head-specific activations to the option subspace. Subsequently, we operationalize the distinction between *Steadfast* and *Wavering* behaviors through entropy time-series analysis. Finally, we outline a selection strategy to aggregate these metrics and identify a compact subset of heads that serve as the most effective targets for intervention.

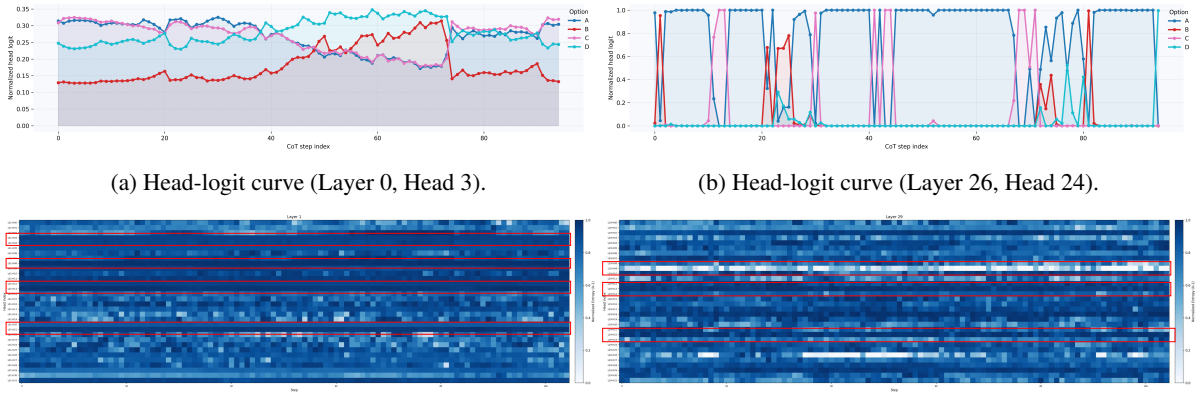
3.1 Stepwise Head-wise Option Readout

Formalized. During CoT generation, we seek a *head-resolved* and *time-aligned* view of the competition among answer options K (e.g., $\{A, B, C, D\}$ for 4-way benchmarks, or $\{A, B, C, D, E\}$ for 5-way benchmarks such as AQUA-RAT). Since attention heads write to the residual stream, their outputs share the vocabulary's latent space. At decoding step t, for head (ℓ, h) with output $o_{\ell,h}^{(t)} \in \mathbb{R}^d$, we apply a logit-lens style unembedding $W_U \in \mathbb{R}^{|V|\times d}$. To isolate the decision signal from noise, we restrict the projection to option tokens:

$$z_{\ell,h}^{(t)}[k] \;=\; (W_U\, o_{\ell,h}^{(t)})[\pi(k)], \quad k \in K, \qquad (1)$$

where $\pi(k)$ maps option label k to its token id (aggregating multi-token labels if necessary). We then normalize these scores over K to obtain a focused option distribution:

$$p_{\ell,h}^{(t)} = \mathrm{softmax}\Big(\beta\, z_{\ell,h}^{(t)}\Big) \in \Delta^{|K|}, \qquad (2)$$



(a) Head-logit curve (Layer 0, Head 3).

(b) Head-logit curve (Layer 26, Head 24).

(c) Entropy heatmap (Layer 1).

(d) Entropy heatmap (Layer 29).

Figure 2: **Visualizing the Entropy Dichotomy.** We map the temporal entropy traces of representative heads during CoT generation on LLaMA-2-7B-chat. Like the attention heads highlighted by the red boxes, heads that remain in deep blue over time are *Wavering Heads*, while heads that stay light-colored or nearly white are *Steadfast Heads*.

where β is an inverse temperature for calibration. This probability vector $p_{\ell,h}^{(t)}$ serves as the per-step decision readout, translating internal activations into an interpretable categorical preference for subsequent entropy analysis.

Extensibility. While we focus on single-answer multiple-choice benchmarks for clarity, the readout is not tied to option letters. More generally, one can define a task-specific token set K that represents the relevant decision symbols and apply the same restricted unembedding and normalization. For instance, in arithmetic evaluation problems, K can be chosen as Arabic numerals and operator symbols (e.g., $\{0, \ldots, 9, +, -, \times, \div, =\}$), and we can measure entropy over this set to track how decisively a head supports the emerging computation or final value. Concrete paths for extending this readout to open-ended generation are outlined in Appendix H.5.

3.2 Entropy Trajectories Tracing

Given the head-wise option distribution $p_{l,h}^{(t)}$, we quantify the uncertainty of head (l, h) at decoding step t via its entropy:

$$\mathcal{H}_{l,h}^{(t)} = -\sum_{k \in \mathcal{K}} p_{l,h}^{(t)}[k] \log p_{l,h}^{(t)}[k]. \tag{3}$$

We use the natural logarithm throughout, so the entropy satisfies $\mathcal{H}_{\ell,h}^{(t)} \in [0, \log |\mathcal{K}|]$.

This yields a temporal trace $\mathbf{H}_{l,h} = (\mathcal{H}_{l,h}^{(1)}, \ldots, \mathcal{H}_{l,h}^{(T)})$ capturing the evolution of option competition.

As visualized in Figure 2, **we observe a robust dichotomy in the *structure* of these traces, motivating a regime-based classification.**

3.3 Probe-Set Aggregation for Stable Head Selection

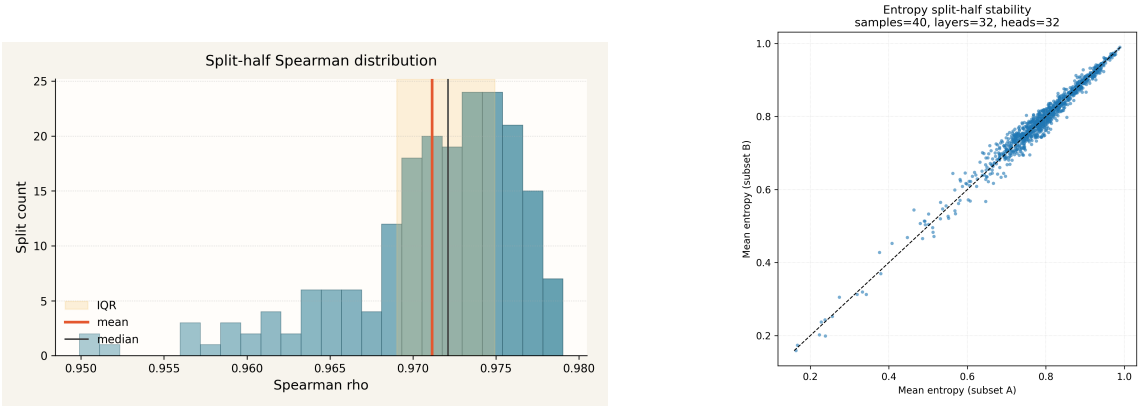
To select a global set of intervention targets, we aggregate head behaviors over a small probe dataset $\mathcal{D}_{probe}$ ($N = 40$). A critical prerequisite for using such a small sample is the statistical stability of the metric.To verify this, we proceed as follows.

Metric Stability. We validate the reliability of our readout via split-half analysis, randomly partitioning the probe set 200 times. **As shown in Figure 3**, the head-wise entropy exhibits exceptional stability. The scatter plot reveals a tight diagonal alignment with a Spearman rank correlation of $\rho = 0.973$ ($p \approx 0$) and a narrow IQR of 0.969–0.975. This confirms that Steadfast and Wavering behaviors are robust intrinsic properties of the heads rather than artifacts of specific samples.

Selection Protocol. Given this stability, we verify the existence of consistent functional heads using the frequency analysis in Appendix C. The structural sparsity we uncovered motivates a robust aggregation strategy to isolate the most canonical instances.We employ the following pipeline:

1. **Per-Sample Candidate Identification:** For each example in the probe set, we identify candidate heads that satisfy strict behavioral thresholds for steadfast or wavering heads.

2. **Global Frequency Ranking:** We aggregate these local candidates and rank them primarily by their recurrence frequency across the entire



(a) Split-half Spearman distribution.

(b) Entropy split-half stability.

Figure 3: Metric Stability Analysis. (a) The distribution of split-half correlations is sharply peaked at $\rho \approx 0.97$. (b) The scatter plot confirms that head rankings are highly consistent across random data splits, validating the robustness of entropy profiling.

probe dataset $\mathcal{D}_{probe}$, utilizing the global time-averaged entropy $\overline{\mathcal{H}}_{l,h}$ as a secondary metric for tie-breaking.

3. **Budget-Constrained Selection:** We instantiate the final intervention sets H_{WH} and H_{SH} by selecting the top-M heads from these ranked lists, ensuring that we target components with the highest behavioral consistency.

We set the budget to $M = 30$. This choice is empirically justified by our subsequent sensitivity analysis (Figure 4), where performance gains saturate at this threshold, indicating that a critical subset of approximately 30 heads is sufficient to capture the primary drivers of decision plasticity. Sensitivity of head selection to the probe-set size N is reported in Appendix H.4.

4 Selective Fine-tuning

Having characterized the internal decision dynamics via Dynamic Entropy Tracing, we now transition from observation to causal intervention. While section 3 established a robust behavioral dichotomy, determining whether these entropy signatures determine actual trainability requires active manipulation. In this section, we employ **Selective Head Fine-Tuning** as a mechanistic probe to test the plasticity of the identified components. And this controlled regime allows us to verify the hypothesis that components sustaining deliberation serve as the primary leverage points for steering reasoning behavior.

4.1 Method

We formulate the intervention as a constrained optimization problem on the Supervised Fine-Tuning (SFT) dataset $\mathcal{D}_{SFT} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i is the prompt and y_i is the target completion ending with an option token in $\mathcal{K}$. Let Θ denote the full model parameters. For each attention head (l, h), let $\theta_{l,h} \subset \Theta$ represent the specific subset of parameters governing its query, key, value, and output projections.

Given a target head set H derived from the selection process (e.g., H_{WH} or H_{SH}), we partition the parameter space into a trainable subset $\Theta_H = \{\theta_{l,h} : (l, h) \in H\}$ and a frozen complement $\Theta_{\setminus H} = \Theta \setminus \Theta_H$.[2] Our **SHFT** objective is to minimize the negative log-likelihood loss $\mathcal{L}$ strictly over Θ_H:

$$\min_{\Theta_H} \mathbb{E}_{(x,y)\sim\mathcal{D}_{SFT}}[\mathcal{L}(f_{\Theta_H,\Theta_{\setminus H}}(x), y)] \quad (4)$$

subject to the constraint that $\Theta_{\setminus H}$ remains fixed at pre-trained values. This protocol ensures that any behavioral shift in the model is causally attributable solely to the plasticity of the selected head set H, isolating the functional role of these components under a rigid parameter budget.

4.2 Intervention Regimes

To isolate the causal contribution of specific head types, we design a comparative framework where

[2]Each selected head contributes its Q, K, V, and O projection slices as trainable parameters; LayerNorm and MLP parameters remain frozen. For LLaMA-2-7B-chat ($d = 4096$, $d_{\text{head}} = 128$), this totals $4 \times (4096 \times 128) \times 30 \approx 62.9$M parameters ($\sim$0.92% of 6.7B). For Qwen3-8B, the trainable fraction is $\sim$0.85%.

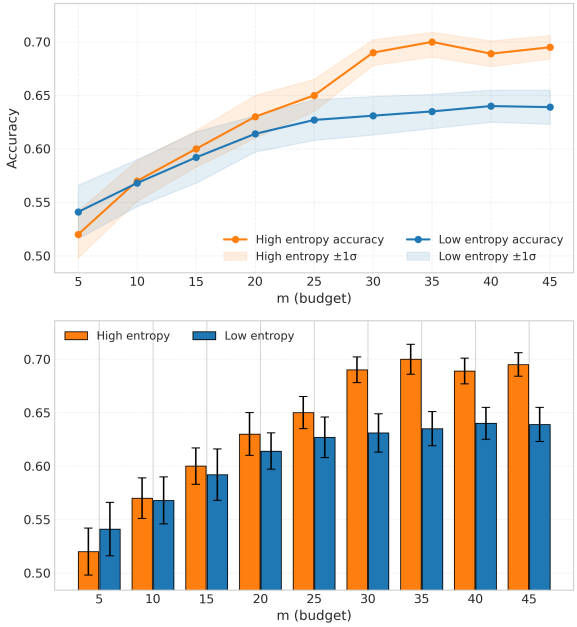


Figure 4: Sensitivity analysis of the head budget M. The Wavering Head strategy (orange) shows a performance ascent that saturates around $M = 30$, empirically justifying our budget choice. Conversely, the Steadfast strategy (blue) exhibits essentially zero sensitivity to budget increases, confirming that these components lack adaptation capacity regardless of the parameter allowance.

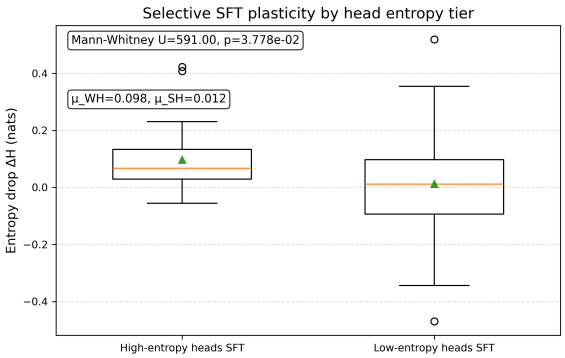


Figure 5: We visualize the entropy reduction $\Delta\mathcal{H} = \mathcal{H}_{\text{pre}} - \mathcal{H}_{\text{post}}$ for the selected heads, where each point corresponds to a single attention head. We compare WAVERING HEADS (WH) against STEADFAST HEADS (SH) using a *two-sided* Mann–Whitney U test (single planned comparison; no multiple-comparison correction). Effect sizes indicate a moderate distributional shift: rank-biserial correlation $|r_{\text{rb}}| = 0.313$ and Cliff's $\delta = 0.354$.

the number of trainable parameters is strictly held constant across all head-only interventions. We fix the head budget to $M = 30$ and evaluate **SHFT-Wavering** approach alongside **Full SFT** (as reference), **SHFT-Steadfast**, **SHFT-Random**, and **LoRA**. All regimes share identical optimization hyperparameters, prompt formats, and data ordering to ensure fair comparison. This design constitutes a direct mechanistic test: if decision plasticity is indeed concentrated in components that sustain uncertainty, **SHFT-Wavering** should approximate Full SFT performance, whereas **SHFT-Steadfast** will yield significantly lower gains. We present the detailed empirical results in Section 5.2.

4.3 What Makes Wavering Heads Effective?

The empirical superiority of tuning Wavering Heads calls for a mechanistic account grounded in the optimization signal(Appendix D). In our setting, the fine-tuning objective is the standard negative log-likelihood over the target completion, whose final line ends with an option token in K. This implies that the relevant training signal is governed by the cross-entropy error term.

At the answer token, the loss gradient over option logits is $p - e_y$, which becomes small only when the model is *confident and correct* ($p[y] \approx 1$), not merely when entropy is low. A low-entropy but wrong prediction can still yield large gradients, so "saturation" requires an alignment condition. Under a frozen backbone, a head's trainability depends on both the error signal $(p - e_y)$ and a local coupling $A_{\ell,h}$ that maps that head's readout to the option logits. This exposes two bottlenecks: error saturation and weak/misaligned coupling. Steadfast heads are hard to improve when they are already correctly-saturated on most fine-tuning samples. Wavering heads more often stay in small-margin regimes, providing non-trivial supervision signals. Although we do not explicitly estimate $A_{\ell,h}$, SHFT acts as a causal check: gains from updating only selected heads imply their couplings are sufficiently usable.

We also validate this theoretical proposition by analyzing the distributional shift of head entropy after fine-tuning. Figure 5 visualizes the entropy drop $\Delta\mathcal{H} = \mathcal{H}_{pre} - \mathcal{H}_{post}$ for the selected intervention targets. The results reveal a stark contrast in plasticity. Wavering Heads (left boxplot) exhibit a significant entropy reduction with a mean drop of $\mu_{WH} = 0.098$. This positive shift confirms that these heads effectively absorb the gradient updates to resolve their uncertainty. In contrast, Steadfast Heads (right boxplot) show a negligible change ($\mu_{SH} = 0.012$), statistically distinct from Wavering Heads ($p < 0.05$, Mann-Whitney U test).

Method	CoT-Collection			AQUA-RAT			arc-cot		
	Rec	F1	Acc	Rec	F1	Acc	Rec	F1	Acc
LLaMA-2-7B-chat	0.467	0.423	0.542	0.239	0.151	0.204	0.435	0.423	0.440
+ Full SFT	0.594	0.574	0.709	0.272	0.274	0.275	0.566	0.567	0.570
+ LoRA	0.526	0.539	0.672	0.179	0.197	0.188	0.573	**0.611**	**0.591**
+ Random-head	0.524	0.525	0.664	0.226	0.240	0.229	0.550	0.551	0.556
+ SH-only	0.542	0.545	0.659	0.240	**0.267**	0.244	0.475	0.476	0.480
+ WH-only	**0.577**	**0.578**	**0.696**	**0.245**	0.253	**0.259**	**0.581**	0.582	0.590
Qwen3-8B	0.594	0.599	0.793	0.519	0.516	0.531	0.716	0.745	0.710
+ Full SFT	0.693	0.668	0.829	0.564	0.563	0.555	0.802	0.794	0.790
+ LoRA	0.701	0.692	**0.857**	0.608	**0.607**	**0.614**	0.802	0.801	0.790
+ Random-head	0.713	0.689	0.845	0.593	0.590	0.599	0.804	0.797	0.793
+ SH-only	0.693	0.680	0.822	0.600	0.590	0.594	0.782	0.773	0.770
+ WH-only	**0.705**	**0.698**	0.849	**0.614**	0.603	0.612	**0.810**	**0.804**	**0.800**

Table 1: Results on three datasets. Columns report Recall / F1 / Accuracy for each fine-tuning regime under each base model.

This corroborates our gradient vanishing analysis: Steadfast Heads act as fixed points resistant to modification, whereas Wavering Heads serve as the plastic components that drive model adaptation.

5 Experiments

5.1 Setup

We evaluate our approach on two open-source LLMs: **LLaMA-2-7B-chat** (Zhao et al., 2021) and **Qwen3-8B** (Yang et al., 2025). Experiments are conducted on three single-answer reasoning benchmarks: **CoT-Collection** (Kim et al., 2023) (from which we filtered the single-choice subset), **AQUA-RAT** (Ling et al., 2017), and **arc-cot**[3].

Comprehensive training configurations and hyperparameters are detailed in the Appendix E. We report accuracy, recall, and F1 score on the final answer token as the primary metrics.

5.2 Main Results: The Plasticity Hierarchy

Table 1 presents the comparative results across two base models and three datasets. We observe a consistent performance hierarchy across all experimental settings: **SHFT-Wavering** $\approx$ **Full SFT** $>$ **LoRA** $>$ **Random** $>$ **SHFT-Steadfast**.

[3]https://huggingface.co/datasets/Locutusque/arc-cot

Recovering Full SFT Performance. Despite updating only 30 attention heads (freezing $> 99\%$ of parameters), SHFT-Wavering effectively closes the gap with, and in some cases surpasses, full-parameter fine-tuning. For instance, on **CoT-Collection** with LLaMA-2-7B-chat, Full SFT achieves an accuracy of 70.9%. Our method reaches 69.6%, recovering nearly all performance gains. Meanwhile, on Qwen models, tuning Wavering Heads modestly exceeds Full SFT on several benchmarks (e.g., 84.9% vs. 82.9% on Qwen3-8B). We attribute this phenomenon to reduced over-parameterization: by restricting updates to non-essential components (like Steadfast Heads), SHFT acts as a form of regularization, preventing the model from overfitting to spurious statistics in the training data and thereby preserving more robust reasoning logic (Zaken et al., 2022).

The Cost of Certitude. Conversely, SHFT-Steadfast yields much less gains, often underperforming even the Random baseline (e.g., 65.9% vs. 66.4% on LLaMA-2-7B-chat, CoT-Collection). This result is mechanistically significant: it confirms that Steadfast Heads, having already collapsed their option distribution, offer minimal leverage for steering. From an optimization perspective, these heads have prematurely minimized local uncertainty, **effectively acting as rigid anchors that**

Method	Wikitext			Emotion		
	BLEU-4	ROUGE-L	PPL	Rec	F1	Acc
LLaMA-2-7B-chat	0.020	0.153	13.496	0.302	0.195	0.228
+ Full SFT	0.012	0.122	47.550	0.326	0.214	0.238
+ WH-only	**0.020**	**0.151**	13.058	0.329	0.208	0.246
+ LoRA	0.016	0.135	**12.651**	**0.367**	**0.263**	**0.281**
Qwen3-8B	0.014	0.134	22.808	0.488	0.467	0.531
+ Full SFT	0.016	0.135	60.191	0.460	**0.457**	**0.532**
+ WH-only	0.016	0.136	18.837	**0.482**	0.453	0.515
+ LoRA	**0.018**	**0.138**	**17.556**	0.459	0.435	0.482

Table 2: Generalization performance on out-of-distribution datasets (Wikitext and Emotion). While **Full SFT** suffers from significant degradation in general language modeling capabilities (indicated by high PPL on Wikitext), our **WH-only** method effectively preserves the model's generalization ability, achieving low perplexity comparable to LoRA while maintaining competitive classification performance.

are difficult to shift. Consequently, they generate negligible error signals during backpropagation, lacking the necessary plasticity to absorb new supervision. Intervening on these "committed" components is thus computationally inefficient.

Comparison with LoRA. SHFT-Wavering also remains competitive with or superior to LoRA in several settings (e.g., 69.6% vs. 67.2% on LLaMA-2-7B-chat), despite the structural simplicity of merely unfreezing existing parameters rather than introducing new modules. Overall, these findings empirically support our central hypothesis: effective adaptation knobs are defined by their process-level uncertainty. Seed variance and random-head draw variance are reported in Appendices H.2 and H.3; cross-task transfer experiments supporting the intrinsic-head-property interpretation are reported in Appendix H.1.

5.3 Generalization and Robustness Analysis

A critical risk in task-specific fine-tuning is catastrophic forgetting, where optimizing for the target reasoning benchmark degrades the model's general language capabilities (Kotha et al., 2024). To assess this, we evaluate our fine-tuned models on two out-of-distribution (OOD) tasks: **Wikitext** (Merity et al., 2017) (to measure generation quality and perplexity) and **Emotion** (Saravia et al., 2018) (to measure classification transfer), as shown in Table 2.

Mitigating Distributional Drift. The results highlight a significant robustness advantage for our approach. As detailed in Table 2, **Full SFT** induces severe distributional drift. On LLaMA-2-7B-chat, the perplexity (PPL) on Wikitext spikes to 47.55, and on Qwen3-8B it reaches 60.19, indicating a substantial degradation in generation fluency. In sharp contrast, **WH-only** maintains low perplexity (13.06 for LLaMA-2-7B-chat, 18.84 for Qwen3-8B), comparable to the **LoRA** baseline. Furthermore, WH-only achieves higher BLEU-4 scores than Full SFT on LLaMA-2-7B-chat (0.020 vs. 0.012), confirming that our method effectively improves reasoning performance without compromising the model's fundamental linguistic backbone.

Task Transfer Capabilities. On the Emotion classification task, **WH-only** demonstrates competitive transfer performance. For Qwen3-8B, our method achieves an F1 score of 0.453, nearly matching Full SFT (0.457) and outperforming LoRA (0.435). On LLaMA-2-7B-chat, it remains comparable to Full SFT (0.208 vs. 0.214). This indicates that the behavioral changes induced by tuning Wavering Heads are not merely overfitting to the CoT format, but reflect a controlled adaptation that preserves broader utility.

Overall, SHFT acts as a safety-aligned, surgical intervention: it corrects specific reasoning behaviors while minimizing the unintended side effects.

6 Conclusion

Our findings translate mechanistic observables into a principled criterion for selective fine-tuning. By proposing **Dynamic Entropy Tracing**, we validate the **Uncertainty-Plasticity Hypothesis**, showing adaptation capacity concentrates in *Wavering Heads*. This occurs because confident heads act as fixed points resistant to updates, whereas uncertain heads drive optimization. Consequently, our **Selective Head Fine-Tuning** recovers Full SFT performance by updating fewer than 1% of parameters, mitigating catastrophic forgetting. Future work may extend this pipeline to larger reasoning-optimized architectures. Ultimately, this confirms that robust steering requires targeting interfaces where decision formation remains active.

Limitations

Our experiments explore a representative but not exhaustive set of model sizes, training seeds, and hyperparameter configurations, so some quantitative results may vary under different tuning choices. We also do not aim to optimize absolute downstream performance; instead, we prioritize controlled comparisons that isolate the effect of head selection. For simplicity, we use a fixed prompting and option-labeling convention throughout, and we do not systematically study alternative templates or label verbalizations. In addition, our analysis emphasizes aggregate trends across heads and examples, while more fine-grained case studies (e.g., per-category or per-error-type breakdowns) are left to future work. We report standard metrics and main findings, but do not attempt a comprehensive sweep over auxiliary diagnostics or ablations.

Due to compute constraints, we did not include dedicated reasoning-optimized models (e.g., Qwen Thinking variants) in our experimental suite. Extending the same screening and selective fine-tuning pipeline to such reasoning models, along with broader scaling studies, is a natural next step that we plan to pursue when additional resources are available.

Ethics Statement

This paper explores the internal decision dynamics of LLMs during CoT reasoning and introduces a mechanism-aware selective fine-tuning strategy to improve adaptivity. The datasets used in our experiments (CoT-Collection, AQUA-RAT, ARC-CoT) are established public benchmarks, and we have reviewed the samples presented in this paper to ensure they do not contain personally identifiable information (PII) or offensive content. All use of existing artifacts, including the LLaMA-2 and Qwen model families, is strictly consistent with their respective research licenses and intended use policies.

The mechanistic insights uncovered in this work, particularly the identification of "Wavering Heads" as the locus of plasticity, provide a pathway for more efficient and safety-aligned model steering. We acknowledge, however, that the ability to surgically modify reasoning logic with minimal parameter updates could potentially be misused to bypass safety guardrails or inject malicious behaviors. Despite this, we believe that exposing these internal mechanisms is a prerequisite for building robust, controllable, and transparent AI systems. We advocate for the responsible disclosure and monitoring of such targeted intervention techniques.

Acknowledgments

This work is supported by the grants from the National Natural Science Foundation of China (No.62376130). The authors would like to thank the organizers of ACL 2026 and the reviewers for their helpful suggestions.

References

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.

Hang Chen, Jiaying Zhu, Xinyu Yang, and Wenya Wang. 2025. Skill path: Unveiling language skills from circuit graphs. *Preprint*, arXiv:2410.01334.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Ching Fang and Samuel Marks. 2025. Unsupervised decoding of encoded reasoning using language model interpretability. *Preprint*, arXiv:2512.01222.

Demi Guo, Alexander M. Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*

the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 4884–4896. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12685–12708. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *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*.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2024. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamile Lukosiute, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *CoRR*, abs/2307.13702.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 158–167. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. 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*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

Christopher Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3687–3697. Association for Computational Linguistics.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià

Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, and 10 others. 2025. Open problems in mechanistic interpretability. *Trans. Mach. Learn. Res.*, 2025.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. 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*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *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*.

Rongwu Xu, Zehan Qi, and Wei Xu. 2024. Preemptive answer "attacks" on chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14708–14726. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1–9. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Datasets Details

We evaluate on three single-answer reasoning benchmarks: CoT-Collection (filtered to a single-choice subset), AQUA-RAT, and arc-cot. These datasets encompass diverse reasoning modalities, ranging from commonsense and scientific inference to multi-step arithmetic.

A.1 CoT-Collection (kaist-ai/CoT-Collection)

CoT-Collection is a large-scale CoT fine-tuning dataset designed to induce explicit reasoning traces in language models. It contains 1.84M CoT-augmented instances spanning 1,060 tasks sourced from the FLAN collection, covering categories such as NLI, QA, science, arithmetic, commonsense, and multiple-choice QA. Each example includes an input prompt (`source`), a reasoning trace (`rationale`), and a final target output (`target`), along with a task identifier. In our experiments, we further filter CoT-Collection to the single-choice subset and map it into our standardized multi-choice format.

A.2 AQUA-RAT (deepmind/AQUA-RAT)

AQUA-RAT is a dataset of algebraic word problems paired with human-written rationales for multi-step quantitative reasoning. Each instance provides a question statement (`question`), a list of candidate answers (`options`), a step-by-step explanation (`rationale`), and the gold label (`correct`). We use it as a single-answer reasoning benchmark under its native five-option (A–E) interface; accordingly, $|\mathcal{K}| = 5$ throughout, and all entropy computations use the corresponding $\log |\mathcal{K}|$ upper bound.

A.3 arc-cot (Locutusque/arc-cot)

arc-cot is an augmented ARC-Challenge science QA benchmark paired with CoT style explanations. The underlying ARC dataset consists of non-diagram, typically 4-way multiple-choice science questions, with a Challenge split intended to be harder than an Easy split. arc-cot provides training examples in a lightweight format: a `question` string containing the prompt and answer choices, and an `answer` string containing the correct option letter followed by an explanation. We use it as a single-answer reasoning benchmark under the same four-option interface.

From CoT-Collection:
content:
Zach drove over to Mary's place. She would be his wife soon. She was in China, visiting her parents. Her son Bradley hadn't gone with her. Bradley was a junior in high school. He neither liked nor disliked Zach, even though he had known Zach for three years. Zach was still trying to get along well with Bradley.
When Zach arrived, he asked if Bradley wanted to drive his car. Bradley had a driver's license. Bradley said all right. Zach told him not to drive fast, but that he could drive anywhere he wanted. Bradley got on the road. Zach gave Bradley a few driving tips: Don't drive next to big trucks, because you never know when they might crush you. Don't drive behind trucks filled with things, because you never know when something in the truck will fly out and hit your car.
On their way back, Zach suggested that they stop at the golf course. He wanted to show Bradley how to play golf. Bradley wasn't interested. He preferred his video games. But Bradley soon discovered that golf was fun! He hit a lot of balls. Zach told him that he was doing well. The next day Bradley, for the first time ever, called Zach. He had a few blisters on his hands. Zach said that usually happened. Then Bradley asked if Zach would come next Saturday so they could take a drive and hit golf balls again. Zach said, of course, and felt happy.

At first, Bradley liked ___ better than golf.
A) driving
B) video games
C) his lessons
D) a driver's license

rationable:
The context mentions that, before spending the day with Zach and getting a chance to try golfing for himself, Bradley preferred video games over playing golf. This implies that his interest in video games is greater than his interest in learn-

ing how to play golf or other activities such as driving. Therefore the most likely answer is B - "video games".

ground-truth: B

From AQUA-RAT:
content:
Two friends plan to walk along a 43-km trail, starting at opposite ends of the trail at the same time. If Friend P's rate is 15% faster than Friend Q's, how many kilometers will Friend P have walked when they pass each other?

A) 21
B) 21.5
C) 22
D) 22.5
E) 23

rationable:
If Q complete x kilometers, then P completes 1.15x kilometers. $x + 1.15x = 43$. $2.15x = 43$. $x = 43/2.15 = 20$. Then P will have have walked $1.15 \times 20 = 23$ km. The answer is E.

ground-truth: E

From ARC:
content:
Sheep Challenge Evolution A species of sheep lives on the Scottish island of Hirta. Hirta has plenty of grass for the sheep to eat, and no natural predators of the sheep live on the island. Some sheep have dark wool, and some have light wool. The sheep with dark wool tend to be much larger and stronger than the sheep with light wool. However, researchers have found that in the last 20 years, the sheep with light wool have increased in number. Usually, strength and size help animals survive and breed. A new sheep predator is introduced to the island. Which change to the sheep population will most likely occur first?

A. a decrease in the population of sheep with dark wool
B. a decrease in the population of sheep with light wool
C. an increase in the population of sheep with dark wool
D. an increase in the population of sheep with light wool.

rationable:
1. **Predator Introduction:** A new predator is introduced to the island.
2. **Selective Pressure:** The predator poses a threat to the sheep population.
3. **Trait Advantage:** Sheep with dark wool are larger and stronger, giving them an advantage in defending themselves against the predator.
4. **Survival and Breeding:** Sheep with dark wool are more likely to survive and successfully breed, passing on their advantageous trait.
5. **Population Decrease:** As a result of the selective pressure, the population of sheep with light wool, which are less equipped to defend themselves, will likely decrease.

ground-truth: B

B Prompts

We fine-tune all models using a unified CoT prompting scheme adapted for multiple-choice questions with K options (e.g., A–D or A–E). To ensure rigorous comparison, the core instructional content remains identical across all models, differing only in the specific chat template (control tokens) required by the respective tokenizers.

B.1 Unified Instruction Content

The core content of our prompt consists of a fixed system instruction and a template-based user query, defined as follows:

Unified Prompt Content

{System Message}: You are a helpful reasoning assistant. You must perform explicit step-by-step reasoning before giving the final answer.

{User Message}: {Question stem + options}
Please think step by step (chain of thought) to solve this single choice question. After reasoning, output on the last line exactly in the format: `Answer: <OPTION_LETTER>`.

B.2 Chat Templates

We wrap the content defined above into the specific formats required by LLaMA-2-7B-chat and Qwen families. We denote the content blocks from the box above as `{System Message}` and `{User Message}`.

LLaMA-2-7B-chat Chat Template

```
[INST] «SYS»
{System Message}
«/SYS»
{User Message} [/INST]
```

Qwen (ChatML) Template

```
<|im_start|>system
{System Message}<|im_end|>
<|im_start|>user
{User Message}<|im_end|>
<|im_start|>assistant
```

Answer format constraint. We require the *last line* of the model response to be exactly `Answer: <OPTION_LETTER>`, where `<OPTION_LETTER>` corresponds to the valid candidate labels for the specific dataset (e.g., $\in \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}\}$ or $\{\mathtt{A}, \ldots, \mathtt{E}\}$). This constraint is enforced consistently across all datasets and training regimes.

C Frequency Analysis

To validate the robustness of our head selection strategy, we analyze the recurrence frequency of candidate heads across the probe dataset $\mathcal{D}_{probe}$. As illustrated in the frequency distribution plot (Figure 6), the identification of both Steadfast and Wavering Heads exhibits a pronounced **heavy-tailed structure**.

This structural sparsity provides two critical mechanistic insights:

Intrinsic Consistency vs. Transient Noise. The distribution reveals a compact core of heads that satisfy the filtering criteria across a vast majority of probe samples (often appearing in $> 75\%$ of instances). This high recurrence confirms that the observed behaviors (sustained uncertainty or early commitment) are *intrinsic functional attributes* of specific heads, rather than transient artifacts dependent on specific input tokens. Conversely, the long tail consists of heads that exhibit these behaviors only sporadically, suggesting they do not play a reliable role in the global decision dynamics.

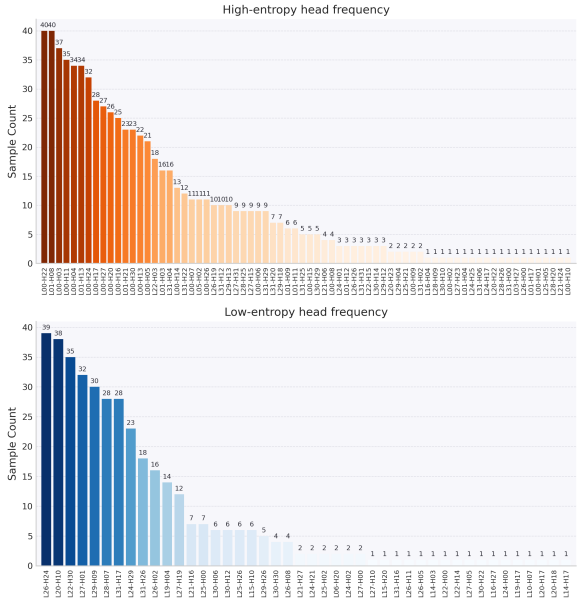


Figure 6: Frequency distribution of candidate heads satisfying the Steadfast (bottom) and Wavering (top) filtering criteria across the 40-sample probe set. The heavy-tailed distribution indicates that a distinct subset of heads consistently exhibits the target behavior, justifying a focused selection budget.

Empirical Grounding for Budget Selection. The sharp drop-off in the frequency curve empirically justifies our strictly constrained intervention budget ($M = 30$). Since the "signal"—represented by consistent participation in the option competition—is highly concentrated in the top tier of the distribution, expanding the selection budget into the long tail would likely yield diminishing returns. This confirms that effectively steering the model requires targeting only the *canonical* subset of heads where decision plasticity is concentrated.

D Theoretical Analysis: Gradient Vanishing.

Let $t^\star$ be the answer-token position and $s \in \mathbb{R}^{|K|}$ the logits over options K, with $p = \mathrm{softmax}(s)$ and gold $y \in K$. For cross-entropy $L = -\log p[y]$,

$$\frac{\partial L}{\partial s} = p - e_y. \tag{5}$$

Crucially, *low entropy alone does not imply small gradients*: a distribution can be low-entropy yet confident on a *wrong* option, yielding a large $\|p - e_y\|$. The relevant "saturation $\Rightarrow$ small gradient" regime is the *correctly-saturated* case, i.e., $p[y] \approx 1$ (equivalently, a large positive margin $m = s_y - \max_{k \neq y} s_k \gg 0$), where $\|p - e_y\| \approx 0$.

To relate this to a head (ℓ, h) under a frozen backbone, we locally linearize the influence of its head-wise readout $z_{\ell,h} \in \mathbb{R}^{|K|}$ at $t^\star$:

$$s \approx s_0 + A_{\ell,h}\, z_{\ell,h}, \tag{6}$$

where $A_{\ell,h}$ is a local coupling matrix. By chain rule,

$$\nabla_{\theta_{\ell,h}} L = \Big(\frac{\partial z_{\ell,h}}{\partial \theta_{\ell,h}}\Big)^\top A_{\ell,h}^\top (p - e_y). \tag{7}$$

Eq. (7) highlights two distinct bottlenecks: (i) *error saturation* when the model is already confidently correct at $t^\star$ ($p[y] \approx 1$), and (ii) *weak/misaligned coupling* ($A_{\ell,h}$) from that head into the option logits. Therefore, interpreting Steadfast (low-entropy) heads as "hard to train" requires the additional condition that their collapse is typically *aligned with the gold label* on the fine-tuning distribution, so that $p - e_y$ is small. Conversely, Wavering (high-entropy) heads more often operate in small-margin regimes where $p - e_y$ is non-trivial, providing usable supervision signals. While our selection is entropy-only and does not explicitly estimate $A_{\ell,h}$, SHFT provides a causal sanity check: if the chosen heads lacked usable coupling, updating only them under a frozen backbone would not yield the consistent gains we observe.

E Training Details

We perform supervised fine-tuning (SFT) with an autoregressive language-modeling objective, where the loss is computed only over the assistant response tokens (the prompt tokens are masked out). Each training example is truncated to a maximum sequence length of 1024 tokens; when truncation is necessary, we keep the suffix to preserve the response portion.

Unless otherwise specified, we train for 10 epochs with a learning rate of 5×10^{-5}. We use a per-device batch size of 1 with gradient accumulation of 8 steps, yielding an effective batch size of $8 \times N$ where N is the number of data-parallel devices. Optimization uses AdamW and a cosine learning-rate schedule with a warmup ratio of 0.03. We log training statistics every 10 update steps. Mixed precision is enabled whenever applicable: we use bfloat16 if supported by the hardware, and otherwise fall back to float16. For memory-constrained settings, we optionally support 8-bit weight loading, and enable gradient checkpointing when needed.

For LLaMA-style experiments, we run epoch-level evaluation on a held-out validation split and retain the checkpoint with the best validation loss (keeping at most one checkpoint). For Qwen-style experiments, we follow the same optimization hyperparameters but omit intermediate validation/checkpointing and save the final model state.

For LoRA-based baselines, we attach low-rank adapters to the attention `q_proj` and `v_proj` modules in all transformer layers, using rank $r = 8$ and scaling $\alpha = 16$ (i.e., $\alpha/r = 2$). We set `bias=none` and use zero LoRA dropout (`lora_dropout=0`). All other optimization hyperparameters follow the default SFT setup described above. When 8-bit weight loading is enabled, we additionally apply the standard k-bit preparation before injecting LoRA adapters.

F Heads Filtering Details

We implement a four-stage pipeline to identify attention heads with characteristic confidence dynamics during step-by-step generation: (i) per-step head-wise logit tracing, (ii) entropy computation, (iii) heatmap visualization, and (iv) dataset-level head selection.

Stage 1: Per-step head-wise logit tracing. For each example x, we perform autoregressive decoding and record, at every generation step $t \in \{1, \ldots, T_x\}$, the contribution of each attention head (ℓ, h) to the logits over the multiple-choice answer options. Let K denote the number of answer options (in our setting $K = 4$ for CoT-Collection/ARC-CoT and $K = 5$ for AQUA-RAT). We denote the recorded head-wise option logits as

$$\mathbf{z}_{\ell,h,t}^{(x)} \in \mathbb{R}^K. \tag{8}$$

Unless otherwise noted, we run deterministic decoding with maximum generation length $T_x \leq 256$, temperature 0, top-$p = 1.0$, and repetition penalty 1.0. To mitigate degenerate copying loops during long generations, we additionally use a sliding-window overlap detector with window size 120

and overlap-ratio threshold 0.9; we allow optional length extension up to a factor of 2 if early stopping is triggered.

Stage 2: Entropy computation and normalization. From head-wise logits, we compute an option distribution with inverse-temperature scaling β:

$$\mathbf{p}_{\ell,h,t}^{(x)} = \mathrm{softmax}(\beta\, \mathbf{z}_{\ell,h,t}^{(x)}), \tag{9}$$

and the corresponding entropy

$$H_{\ell,h,t}^{(x)} = -\sum_{k=1}^{K} p_{\ell,h,t}^{(x)}(k)\, \log p_{\ell,h,t}^{(x)}(k). \tag{10}$$

We set $\beta = 1.0$ throughout unless specified otherwise. For visualization and robust thresholding, we further apply a per-example min–max normalization across all heads and steps:

$$\widetilde{H}_{\ell,h,t}^{(x)} = \mathrm{clip}\left(\frac{H_{\ell,h,t}^{(x)} - H_{\min}^{(x)}}{H_{\max}^{(x)} - H_{\min}^{(x)}},\, 0,\, 1 \right), \tag{11}$$

where $H_{\min}^{(x)} = \min_{\ell,h,t} H_{\ell,h,t}^{(x)}$ and $H_{\max}^{(x)} = \max_{\ell,h,t} H_{\ell,h,t}^{(x)}$. This normalization maps entropy to $[0, 1]$ within each example and is used only for visualization and thresholding; all head ranking and statistics are computed on the original (unnormalized) entropy values.

Stage 3: Heatmap visualization. For each example x, we summarize each head by its time-averaged normalized entropy

$$\overline{H}_{\ell,h}^{(x)} = \frac{1}{T_x} \sum_{t=1}^{T_x} \widetilde{H}_{\ell,h,t}^{(x)}, \tag{12}$$

and render a layer×head heatmap $\overline{\mathbf{H}}^{(x)} \in \mathbb{R}^{L \times H}$. We use a blue colormap and rasterize figures at 200 DPI.

Stage 4: Per-example candidates and dataset-level selection. We first identify candidate heads on each example by detecting persistent *low-entropy* ("white-band") or *high-entropy* ("blue-band") segments in $\widetilde{H}_{\ell,h,t}^{(x)}$. For low-entropy heads, we compute:

$$f_{\ell,h}^{(x)} = \frac{1}{T_x} \sum_{t=1}^{T_x} \mathbb{1}[\widetilde{H}_{\ell,h,t}^{(x)} < \tau_{\text{low}}], \tag{13}$$

$$m_{\ell,h}^{(x)} = \frac{1}{T_x} \sum_{t=1}^{T_x} \widetilde{H}_{\ell,h,t}^{(x)}, \tag{14}$$

and $r_{\ell,h}^{(x)}$ is equal to max consecutive run length of $1[\widetilde{H}_{\ell,h,t}^{(x)} < \tau_{\text{low}}]$. A head is marked as a *low-entropy candidate* for example x if

$$f_{\ell,h}^{(x)} \geq \alpha, \quad r_{\ell,h}^{(x)} \geq c, \quad m_{\ell,h}^{(x)} \leq \mu, \tag{15}$$

with τ_{low}=0.05, α=0.6, c=10, and μ=0.25. (High-entropy candidates are defined analogously by flipping the inequality, using a high-entropy threshold and requiring the mean entropy to be above a given bound.)

Frequency-first ranking and top-30 selection. Given a dataset $\mathcal{D}$ of N examples, we aggregate candidates into a dataset-level score per head. Let $s_{\ell,h}^{(x)} \in \{0, 1\}$ indicate whether head (ℓ, h) is selected as a candidate on example x under the above criteria. We define the *selection frequency*

$$F_{\ell,h} = \frac{1}{N} \sum_{x \in \mathcal{D}} s_{\ell,h}^{(x)}, \tag{16}$$

and the auxiliary *time-averaged entropy*

$$\overline{m}_{\ell,h} = \frac{1}{N} \sum_{x \in \mathcal{D}} m_{\ell,h}^{(x)}. \tag{17}$$

We rank heads primarily by $F_{\ell,h}$ (descending). For ties or near-ties, we break ties using $\overline{m}_{\ell,h}$ (ascending for low-entropy sets, descending for high-entropy sets). Finally, we select the top M=30 heads for downstream selective fine-tuning and analysis.

Practical considerations. The pipeline supports processing either explicit example IDs or contiguous subsets (offset/limit) and allows skipping individual stages for efficiency. To reduce disk usage, we optionally discard intermediate step-level traces after producing per-head entropy time series and visualizations.

Following this procedure, we present a representative set of entropy heatmaps in the appendix to qualitatively illustrate the recurring head-wise patterns and their cross-example consistency.

G Heads Filtering Results

In this section, we provide extended visualizations of the Dynamic Entropy Tracing results to substantiate the quantitative analysis presented in the main text. By mapping the spatiotemporal evolution of head-wise uncertainty, these heatmaps offer qualitative evidence for the functional dichotomy between components that sustain deliberation and those that commit early.

LLaMA-2-7B-chat on CoT-Collection Figure 7 presents representative entropy traces for LLaMA-2-7B-chat. The visualization clearly demarcates the two functional distinct behaviors: **Wavering Heads** manifest as continuous "blue bands" , whereas **Steadfast Heads** appear as "white bands" .

Qwen3-8B on AQUA-RAT Figure 8 extends our analysis to Qwen3-8B on the AQUA-RAT benchmark. Despite differences in model architecture and the quantitative nature of the reasoning task, the entropy signatures remain highly consistent.

H Additional Robustness and Transfer Analyses

This appendix consolidates supplementary robustness, transfer, and sensitivity analyses. All experiments follow the training setup in Appendix E.

H.1 Cross-Task Transfer of Wavering Heads

To assess whether Wavering Heads reflect an intrinsic, task-consistent head property, we evaluate cross-task transfer: select heads on a source dataset, then fine-tune and evaluate on a target dataset. Table 3 reports the pairwise overlap of the top-30 Wavering Head sets across benchmarks; Tables 4 and 5 report transfer accuracy for LLaMA-2-7B-chat and Qwen3-8B, respectively. Cross-task transfer incurs only modest degradation (mean $\Delta = -0.012$ for LLaMA, -0.008 for Qwen) while consistently outperforming SH-only, and the degradation correlates with head-set mismatch ($r \approx -0.87$), supporting the interpretation of Wavering Heads as an intrinsic architectural property.

Source ↔ Target	LLaMA-2	Qwen3
CoT ↔ AQUA	24/30 (80.0%)	18/30 (60.0%)
CoT ↔ ARC	19/30 (63.3%)	20/30 (66.7%)
AQUA ↔ ARC	18/30 (60.0%)	16/30 (53.3%)
Three-way core	17/30 (56.7%)	16/30 (53.3%)

Table 3: Pairwise overlap of top-30 Wavering Head sets across benchmarks.

H.2 Seed Variance

Table 6 reports accuracy over 3 random seeds on LLaMA-2-7B-chat. On CoT-Collection and ARC-CoT, the WH–SH gap (3.7 and 10.9 points) far exceeds the standard deviations of both WH-only and SH-only ($\leq$ 1.0 point). On AQUA-RAT the gap is narrower (1.5 points) but still exceeds individual standard deviations. WH-only also exhibits the smallest variance across benchmarks, consistent with targeting a mechanistically stable subset.

Head Source	CoT-Coll.	AQUA-RAT	ARC-CoT
CoT-Collection (native)	0.696	0.252	0.576
AQUA-RAT	0.688	0.259	0.571
ARC-CoT	0.679	0.246	0.590
SH-only (ref.)	0.659	0.244	0.480
Random-head (ref.)	0.664	0.229	0.556

Table 4: Cross-task transfer accuracy on LLaMA-2-7B-chat.

Head Source	CoT-Coll.	AQUA-RAT	ARC-CoT
CoT-Collection (native)	0.849	0.604	0.795
AQUA-RAT	0.842	0.612	0.790
ARC-CoT	0.845	0.600	0.800
SH-only (ref.)	0.822	0.594	0.770
Random-head (ref.)	0.845	0.599	0.793

Table 5: Cross-task transfer accuracy on Qwen3-8B.

Method	CoT-Coll.	AQUA-RAT	ARC-CoT
Full SFT	0.706 ± 0.011	0.271 ± 0.014	0.567 ± 0.009
WH-only	0.693 ± 0.008	0.256 ± 0.011	0.587 ± 0.008
SH-only	0.656 ± 0.010	0.241 ± 0.012	0.478 ± 0.010
Random-head	0.660 ± 0.014	0.226 ± 0.016	0.552 ± 0.013

Table 6: Seed variance (mean $\pm$ std over 3 seeds) on LLaMA-2-7B-chat.

H.3 Random-Head Draw Variance

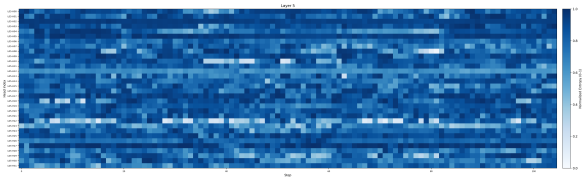
Table 7 reports variance across 3 independent draws of 30 random heads on LLaMA-2-7B-chat. Even the best random draw does not reach WH-only performance on any benchmark.

Statistic	CoT-Coll.	AQUA-RAT	ARC-CoT
Mean $\pm$ std ($n = 3$)	0.662 ± 0.012	0.228 ± 0.010	0.552 ± 0.010

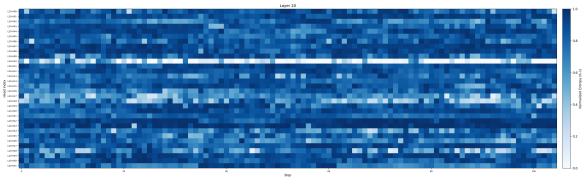
Table 7: Random-head baseline variance (3 independent draws of 30 heads) on LLaMA-2-7B-chat.

H.4 Probe-Set Sensitivity

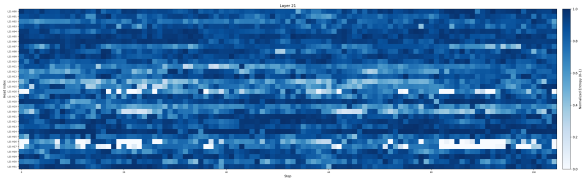
Tables 8 and 9 report the sensitivity of head selection to the probe-set size N. Performance degrades only $\sim$0.5–1 point at $N = 20$ and plateaus by $N = 40$. Even at $N = 20$, WH-only remains well above SH-only (e.g., 0.689 vs. 0.656 on CoT-Collection for LLaMA-2).



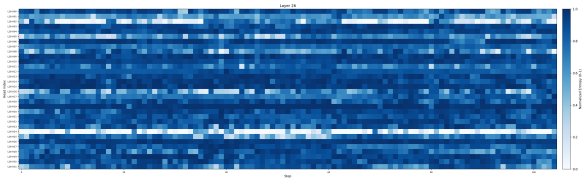
(a) L5-H5 as WH



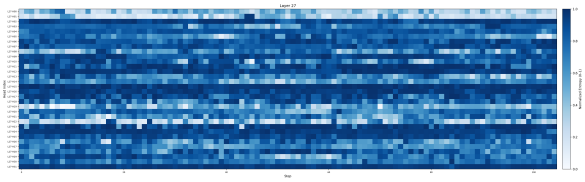
(b) L20-H10 as SH, L20-H23 as WH



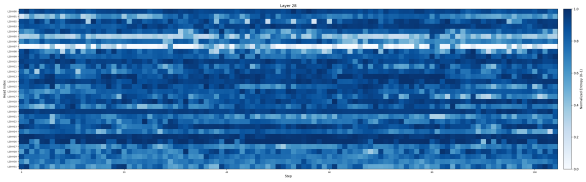
(c) L21-H24 as WH



(d) L26-H24 as SW, L26-H19 as WH



(e) L27-H2, L27-H15, L27-H31 as WH,



(f) L28-H7 as SH, L28-H25 as WH

Figure 7: LLaMA2-7B-chat on CoT-Collection

N	**Overlap w/** N=40	**CoT-Coll.**	**AQUA-RAT**	**ARC-CoT**
20	83.3%	0.689 ± 0.007	0.251 ± 0.009	0.583 ± 0.008
40 (default)	—	0.696 ± 0.004	0.259 ± 0.006	0.590 ± 0.005
60	90.0%	0.698 ± 0.003	0.261 ± 0.005	0.591 ± 0.004

Table 8: Probe-set size sensitivity on LLaMA-2-7B-chat.

H.5 Paths Toward Open-Ended Extensibility

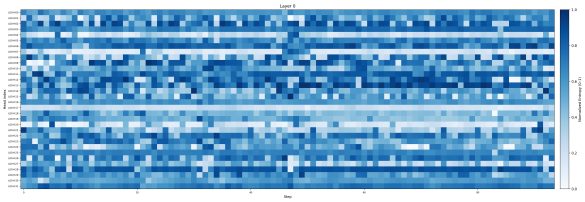
While our main analysis focuses on multiple-choice reasoning, the Dynamic Entropy Tracing framework is not inherently tied to discrete option sets. We outline three concrete paths for extending the readout to open-ended or structured-generation tasks:

(1) Top-k dynamic sets. At each decoding step, $\mathcal{K}$ is adaptively constructed from the model's own top-k predictions. This lets entropy tracing measure head-level distribution sharpness without a pre-specified answer space and is directly applicable to open-ended generation.

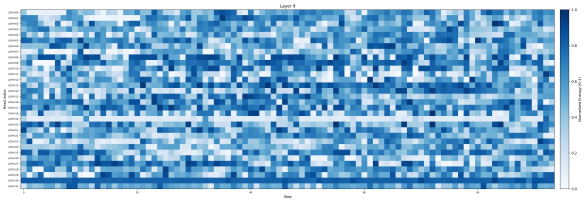
(2) Semantic clustering. $\mathcal{K}$ is defined over curated sets of semantically meaningful token groups (e.g., sentiment-indicative tokens, topic anchors), allowing entropy tracing to capture higher-level semantic decisions rather than literal answer labels.

(3) Contrastive probing. Entropy is computed over tokens representing competing hypotheses in a task-specific manner, capturing deliberation dynamics when the answer space is implicit (e.g., factuality, stance).

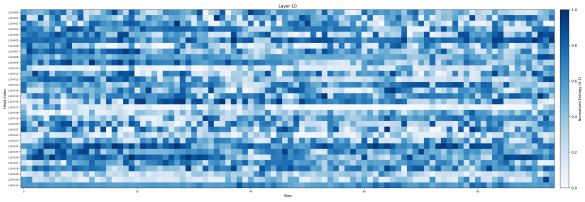
We consider each of these a promising direction for follow-up work, particularly for tasks such as GSM8K, open-domain QA, and long-form reasoning where the relevant decision variable is not a discrete option letter.



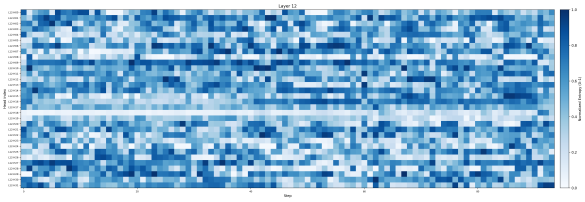
(a) L0-H4, L0-H7, L0-H17 as SH



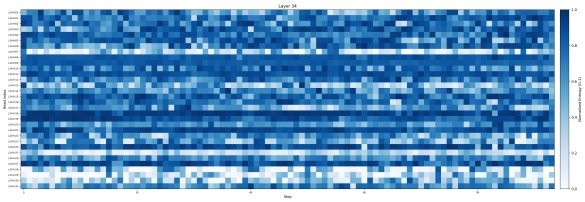
(b) L9-H30 as WH



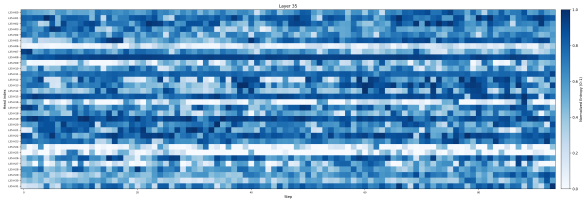
(c) L10-H10 as SH, L10-H31 as WH



(d) L12-H18 as SH



(e) L34-H7, L34-H25 as SH, L34-H8, L34-H9, L34-H18, L34-H21 as WH



(f) L35-H6, L35-16, L35-H24, L25-H25 as SH, L35-H8, L35-H15, L35-H19 as WH

Figure 8: Qwen3-8B on AQUA-RAT

N	**Overlap w/** N=40	**CoT-Coll.**	**AQUA-RAT**	**ARC-CoT**
20	86.7%	0.843 ± 0.006	0.604 ± 0.008	0.794 ± 0.007
40 (default)	—	0.849 ± 0.003	0.612 ± 0.005	0.800 ± 0.004
60	93.3%	0.850 ± 0.002	0.614 ± 0.004	0.801 ± 0.003

Table 9: Probe-set size sensitivity on Qwen3-8B.