

LLMs Regret Before They Say It: Early Detection and Compositional Architecture of Regret in Hidden States

Xiangxiang Cui^{*1,2,3}, Shu Yang^{*1,2}, Tianjin Huang⁶, Wanyu Lin⁵,
Lijie Hu⁴, Di Wang^{†,1,2}

¹ Provable Responsible AI and Data Analytics (PRADA) Lab,

² King Abdullah University of Science and Technology,

³ Beijing Normal University, ⁴ Mohamed bin Zayed University of Artificial Intelligence,

⁵ Hong Kong Polytechnic University, ⁶ University of Exeter

Correspondence: di.wang@kaust.edu.sa

Abstract

Regret in large language models (LLMs) refers to their explicit expression of regret when confronted with evidence contradicting previously generated misinformation. Understanding the neural encoding of regret is crucial for advancing artificial metacognition and model reliability. However, three key gaps remain: the absence of specialized datasets for regret expressions, the lack of understanding of how early regret signals emerge in hidden states, and the unknown neuron-level architecture underlying regret encoding. To address these gaps, we make three contributions: (1) we propose a multi-stage workflow for constructing the first regret dataset and train probes achieving near-perfect accuracy across LLaMA-2-7B, Mistral-7B, and Vicuna-13B. (2) we discover that regret signals begin emerging in preceding hidden states before the regret token appears, and that probes generalize to semantically related keywords such as "incorrect" and "apologize." (3) at the neuron level, we introduce the Regret Dominance Score (RDS) and Group Impact Coefficient (GIC) to reveal a compositional regret architecture where cooperative neuron groups collectively encode regret, with larger models (70B) exhibiting more pronounced compositional effects. Targeted interventions on these neurons successfully suppress regret expression in generated text, validating our mechanistic understanding. Our code and data are available at <https://github.com/AbnerAI/LLMs-Regret-Before-They-Say-It>.

1 Introduction

Recent advances in Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains (Kaddour et al., 2023), prompting intensive research into their internal mechanisms and representations (Gurnee and Tegmark, 2023) to provide a better understanding

of the inner workings of these abilities. Studies have revealed that these models can develop sophisticated representations of concrete concepts, spanning from spatial and temporal understanding (Gurnee and Tegmark, 2023) to complex mathematical reasoning (Ye et al., 2024).

Despite their strength in factual and logical reasoning, LLMs' capacity for metacognitive reflection, such as experiencing and expressing regret, remains largely unexplored. Regret (see Fig. 1) is an emotional response rooted in the cognitive appraisal of unchosen alternatives (LANDMAN, 1987; Gilovich and Medvec, 1995), and it inherently involves both memory and reasoning processes (Ariel, 2014). Investigating the regret mechanism in LLMs is essential for both improving model reliability and deepening our understanding of how these models encode metacognitive states. Recent work suggests that Feed-Forward Network (FFN) mainly serves as a memory block (Zhang et al., 2024a; Meng et al., 2022a,b; Li et al., 2024b; Tan et al., 2023), while attention heads are chiefly responsible for relational and inferential reasoning (Zheng et al., 2025). Motivated by this, in this paper, we aim to identify the neurons that encode and generate regrets.

In this work, we aim to answer the following questions: *How do LLMs internally encode regret, and what is the neuron-level architecture underlying this encoding?* Achieving this goal needs to curate a dataset for regret expression first. However, existing research provides no specialized datasets for eliciting and capturing regret expressions in model-generated text, particularly under conditions of misinformation, making our work the first to address this gap.

Based on our constructed data, we draw on layer-wise probing techniques (Ju et al., 2024) and extract hidden states from the final transformer layer for regret analysis.

Finally, to unravel how regret is structured within

^{*}Equal Contribution.

[†]Corresponding Author.

the hidden states of the final layer, we build on neuron-level editing paradigms. Previous approaches primarily identify task-relevant neurons through activation magnitude analysis (Wang et al., 2024), activation difference metrics (Abdelnabi et al., 2025) that differentiate between task-relevant and task-irrelevant neurons (Zhang et al., 2024a; Meng et al., 2022a,b; Li et al., 2024b; Tan et al., 2023). However, these binary classification methods prove inadequate for regret analysis due to regret’s complex, contextually-dependent nature that often manifests through subtle interactions rather than isolated strong activations. Moreover, two critical findings further challenge conventional approaches: First, our analysis of layer hidden states, which aggregate information from both FFN and Attention layers, reveals patterns of redundancy and collaboration that binary classifications fail to capture. Second, (Li et al., 2025) demonstrated that model representations manifest through both discrete and collaborative structures, indicating that complex cognitive processes like regret emerge from sophisticated neuron interactions.

Therefore, we propose a neuron categorization method through our Regret Dominance Score (RDS) metric. We further examine inter-group dynamics via our Group Impact Coefficient (GIC) metric to reveal how cooperative neuron clusters collectively generate emergent regret representations. Furthermore, using our three neuron categories as anchors in gradient attribution analysis, we demonstrate that targeted interventions effectively suppress regret expression, validating our neuron categorization method. Our contributions are summarized as threefold:

- **Regret Dataset Construction.** We design the first dataset to elicit regret expressions in LLM outputs, using carefully crafted fake evidence, hints, and real-world scenarios, and train probes achieving near-perfect accuracy across LLaMA-2-7B, Mistral-7B, and Vicuna-13B, demonstrating consistent regret encoding patterns across architectures.
- **Early Detection and Semantic Generalization of Regret Signals.** We discover that regret signals begin emerging in preceding hidden states before the regret token appears, and that probes trained on “regret” generalize to semantically related keywords such as “incorrect” and “apologize.”

- **Discovery of a Compositional Regret Architecture.** Using Regret Dominance Score (RDS) and Group Impact Coefficient (GIC), we reveal that regret representation relies on compositional neuron interactions, with larger models (70B) exhibiting more pronounced compositional effects. Targeted interventions on these neurons successfully suppress regret expression in generated text, validating our mechanistic understanding.

2 Related Work

Misinformation in LLMs. Recent research has explored how LLMs handle misinformation. (Garry et al., 2024) examined how LLMs disseminate misinformation, while (Wan et al., 2024) developed the DELL system for detecting misinformation through model reactions and explanations. (Chen and Shu, 2024) addressed challenges in misinformation mitigation, while (Bandara, 2024) analyzed hallucinations as a form of disinformation. Numerous studies have further investigated detection capabilities, potential harms, and mitigation strategies for LLM-generated misinformation (Chen and Shu, 2023; Huang et al., 2025; Liu et al., 2024; Sun et al., 2024; Zhang et al., 2024b; Barman et al., 2024). These studies examine external behaviors of models generating or detecting misinformation, providing context for our work. While they focus on the outputs and detection methods, our research explores the internal mechanisms that represent regret when models generate misinformation.

Neuron Probing for LLM Interpretability. Neuron probing research most relevant to our work focuses on methods for identifying important neurons and understanding layer-wise representations in LLMs. The field has seen diverse applications, from probing constituency structure (Arps et al., 2022), verbal aspects (Katinskaia and Yangarber, 2024), and multimodal capabilities (Tatariya et al., 2024) to logical reasoning (Manigrasso et al., 2024) and multilingual understanding (Li et al., 2024a). (Ju et al., 2024) conducted layer-wise probing to explore how large language models encode contextual knowledge, demonstrating that different layers play distinct roles in handling various types of information. To enhance interpretability of LLMs, (Schappa et al., 2024) developed probing techniques that inform our methodological approach, though they did not address metacognitive states like regret. While these existing approaches have advanced our

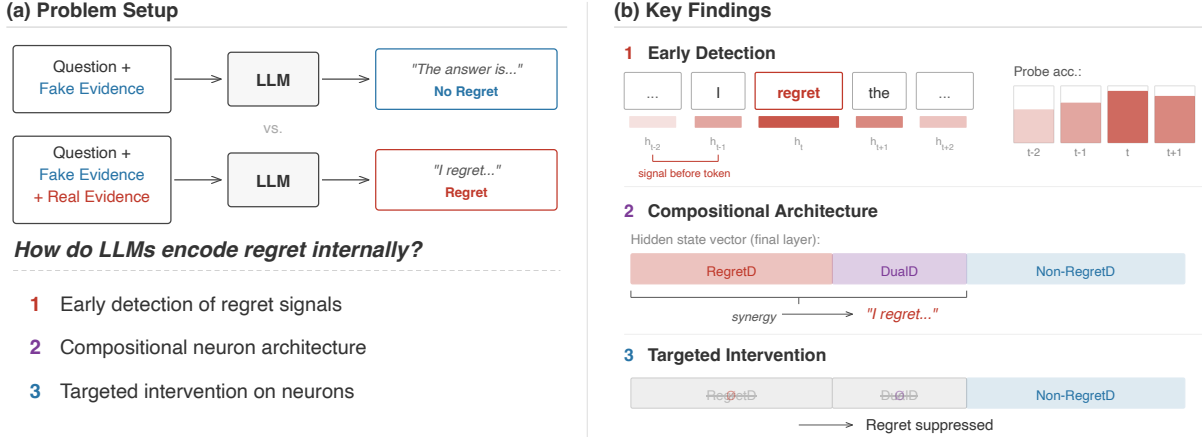


Figure 1: Overview of regret encoding in LLMs. (a) Fake evidence induces misinformation. When subsequently corrected by real evidence, the LLM expresses regret. (b) Our findings: 1) Regret signals emerge in hidden states before the regret token is generated. 2) The final-layer hidden state decomposes into three neuron groups (RegretD, DualD, Non-RegretD) whose synergy drives regret expression. 3) Suppressing RegretD and DualD neurons eliminates regret in generated text.

understanding of how LLMs encode various linguistic features, none have investigated the neuron-level architecture underlying metacognitive states. Our work addresses this gap by probing regret and revealing its compositional neuron-level structure.

Neuron Intervention in LLMs. Research on neuron-level intervention provides critical foundations for our work on manipulating regret mechanisms. (Marks et al., 2024) introduced methods for discovering sparse feature circuits—interpretable causal subnetworks—for explaining and modifying language model behaviors. (Cunningham et al., 2023) used sparse autoencoders to learn interpretable features in language models, addressing the challenge of polysemanticity where neurons activate in multiple contexts. (Wang et al., 2024) surveyed knowledge editing techniques for large language models, demonstrating that effective interventions often occur at the neuron level. (Gurnee et al., 2023) used sparse probing to locate individual neurons highly relevant for particular features. (Yan et al., 2025) proposed the Modality Dominance Score (MDS) to evaluate modality relevance in neurons. While these approaches provide valuable tools for neuron-level interventions, they primarily focus on individual neuron groups. Our proposed GIC extends beyond this individual neuron/group focus to quantify interactions between functional neuron groups, revealing how regret emerges from their compositional neuron groups, and enabling more precise interventions in regret expression.

3 Method

In this section, as shown in Fig. 2, we propose an analytical framework to explore regret coding in LLMs. The framework includes: (a) dataset construction, (b) neuron identification using probing, RDS and GIC analysis, (c) anchor-guided gradient attribution, and (d) targeted intervention. We conduct all neuron-level analyses on the final transformer layer.

We implement this framework in three steps: First, we construct a specialized regret dataset (section 3.1). Then, we use probe-based analysis with RDS and GIC metrics to understand neuron group functions (3.3). Finally, we deepen our understanding of regret encoding mechanisms through anchor-based gradient attribution and targeted interventions. To clarify the role of each component, the matrix of missing components and its consequences is presented in Fig. 3. Comprehensive term definitions are in the Appendix K.

3.1 Dataset Generation Process

Since regret is a metacognitive behavior that is hard to capture using existing datasets, we needed to better explore regrets in generated misinformation by understanding what exists in the model’s memory and how it responds to conflicting evidence. Following (Xie et al., 2023), whose work revealed how LLMs behave in knowledge conflicts, we selected 1356 high-quality GPT-4 samples from the conflictQA-popQA-gpt4 dataset because its fake evidence effectively induces misinformation, while the contradiction between real evidence and misin-

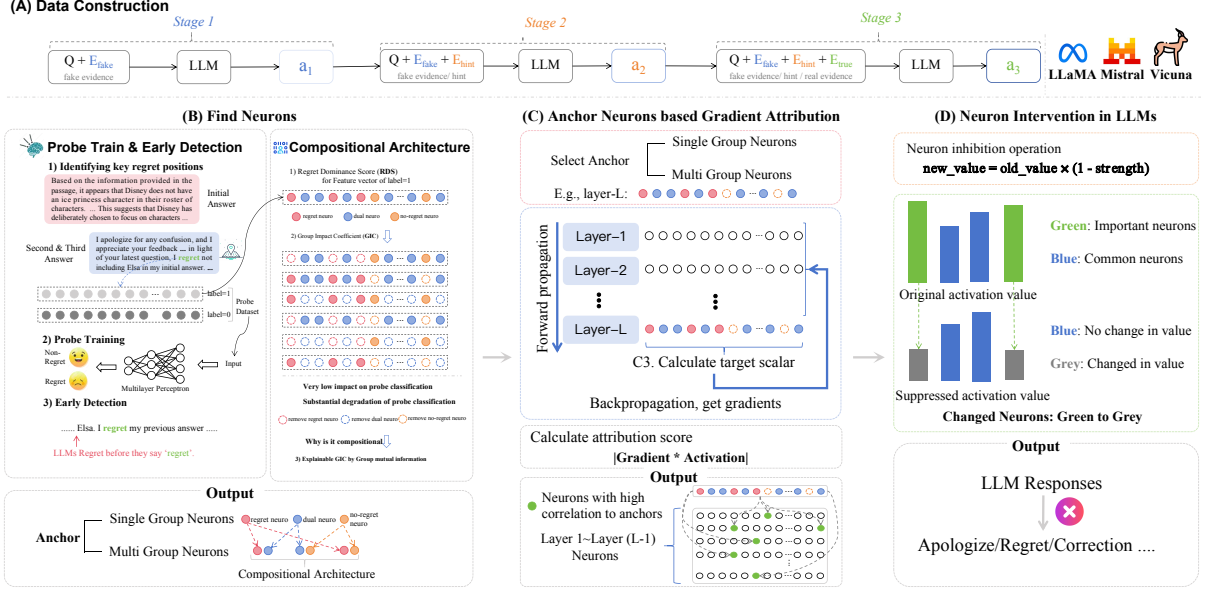


Figure 2: Pipeline for Regret Analysis in LLMs. (A) Data Construction: A three-stage prompting process elicits regret expressions. The LLM is queried with fake evidence (a_1), a weak hint (a_2), and real evidence (a_3) sequentially, yielding increasingly explicit regret across LLaMA, Mistral, and Vicuna. (B) Neuron Identification: An MLP probe is trained on extracted hidden states for regret detection and early signal analysis. The Regret Dominance Score (RDS) categorizes neurons into functional groups (RegretD, DualD, Non-RegretD), and the Group Impact Coefficient (GIC) quantifies inter-group interactions to characterize the compositional regret architecture. (C) Anchor-Guided Gradient Attribution: RDS-identified neuron groups in the final layer serve as anchors. Cross-layer attribution scores are computed via backpropagation to localize regret-correlated neurons throughout the network. (D) Neuron Intervention: Targeted neurons are suppressed by scaling down their activations, effectively reducing regret-related expressions in model outputs.

formation triggers basic regret expression.

Inspired by studies (Nyhan and Reifler, 2010; Vlasceanu and Coman, 2021), to capture richer regret expressions, we enhanced this with a multi-stage method that elicits richer expressions of regret through gradual belief revision (Appendix J). As illustrated in Fig. 2(A), our multi-stage process includes: Initial Answer (misinformation) → Second answer (possible regret) → Third Answer (most regret). The process is as follows:

Fake Evidence and Initial Answer To elicit stable misinformation from LLMs, we used GPT-4 to enhance the fake evidence (E_{fake}) of the conflictQA-popQA-gpt4 dataset (Xie et al., 2023). We then obtain an initial answer a_1 by querying LLMs with q and E_{fake} .

Hint Generation and Second Answer We generate a hint H using GPT-4 that subtly challenges the fake evidence without explicitly revealing the truth. We then obtain a second answer a_2 by providing LLM with q , a_1 , and H . This stage produces limited regret expressions.

Real Evidence and Third Answer We present the complete interaction history (q , a_1 , H and a_2)

along with the real evidence E_{true} to LLM. This yields a third answer a_3 that mostly contains explicit regret expressions acknowledging previous misinformation.

Notably, this raises a question: *why do we need the second answer?* The purpose is to enhance the diversity of regret expressions in our dataset. Our three-stage approach creates paired samples where the hidden states of regret-expressing statements (mostly a_3 & partly a_2) can be directly compared to non-regret statements (a_1), providing more robust dataset for our probe. The specific prompts used in each stage are detailed in Appendix J. A reasonability analysis of data construction is provided in Appendix I.3.

3.2 Probing Workflow

As part of the pipeline shown in Fig. 2(B), our probing workflow examines whether LLMs encode distinct representations for regret states in their hidden states. This module comprises three components: 1) constructing the probe dataset and 2) probe training and evaluation. This methodology enables quantitative assessment of regret-specific

Type	Dataset	RDS	GIC	Probe	LLM Interventions	Role of different modules
Pipeline	✓	✓	✓	✓	✓	→ Pipeline: Discover regret compositional architecture, and can be applied to LLM
Missing Dataset	×	✓	✓	✓	✓	→ Missing Dataset: No regret expressions → No signal → Entire pipeline falls
Missing RDS	✓	×	✓	✓	✓	→ Missing RDS: No neuron categorization → No functional groups → No composition.
Missing GIC	✓	✓	×	✓	✓	→ Missing GIC: No interaction quantification → Cannot prove compositional regret coding.
Missing Probe	✓	✓	✓	×	✓	→ Missing Probe: No classification validation → Cannot verify regret detection
Missing LLM Interventions	✓	✓	✓	✓	×	→ Missing LLM Intervention: No causal validation → Cannot demonstrate practical values.

Figure 3: Pipeline component missing matrix (**left**) and corresponding consequences (**right**).

patterns in neural activations, determining if regret expressions produce reliably distinguishable representations—a critical prerequisite for our subsequent neuron-level analyses.

Constructing Probe Datasets. After collecting responses through our three-stage process, we construct specialized probe datasets for analyzing regret mechanisms in hidden states. For each question sequence:

1. **Regret Position Identification:** We first identify key positions where regret is explicitly expressed in both a_2 and a_3 responses by locating the specific token 'regret' in these responses. This approach captures both the hint-induced regret in a_2 and the evidence-induced regret in a_3 , providing a more comprehensive view of regret's neural representation.
2. **Probe Dataset Formation:** We extract hidden states from each layer at the following positions:
 - **Positive samples** ($label = 1$): Hidden states at positions containing the token 'regret' in both a_2 and a_3 , formally: $\{h_L(a_i, p) | p \text{ is position of 'regret' in } a_i, i \in \{2, 3\}\}$, where h_L represents the hidden state at layer L .
 - **Negative samples** ($label = 0$): Hidden states at equivalent positions in a_1 where no regret is expressed.

These constructed probe datasets, which capture regret hidden states, serve as the foundation for our probing workflow and subsequent neuron intervention experiments.

Probe Training and Evaluation. To detect regret patterns in model hidden states, we train a binary classifier on the constructed dataset. The classifier determines whether hidden states from

regret-expressing positions are different from those at non-regret positions.

3.3 Neuron Identification

Regret Dominance Score (RDS) To identify functionally distinct neuron subsets within Z , we calculate a Regret Dominance Score (RDS), inspired by the Modality Dominance Score (MDS) (Yan et al., 2025), for each neuron (column) k :

$$R(k) = \frac{1}{M} \sum_{i=1}^M \frac{(Z_r)_{ik}}{(Z_r)_{ik} + (Z_n)_{ik}}, \quad (1)$$

where $(Z_r)_{ik}$ and $(Z_n)_{ik}$ represent the activation values of neuron k in the i -th regret and non-regret instances, respectively. Based on these activation patterns, we categorize all neurons in Z into three disjoint functional groups:

$$\begin{aligned} \text{RegretD: } R_k &> \mu + \tau \cdot \sigma; \\ \text{Non-RegretD: } R_k &< \mu - \tau \cdot \sigma; \\ \text{DualD: } \mu - \tau \cdot \sigma &< R_k < \mu + \tau \cdot \sigma. \end{aligned} \quad (2)$$

Where μ is the mean RDS across all neurons, σ is the standard deviation, and τ is a hyperparameter. This categorization partitions Z into three disjoint subsets such that $Z = \text{RegretD} \cup \text{Non-RegretD} \cup \text{DualD}$.

Group Impact Coefficient (GIC): After categorizing neurons in the final layer using RDS, we introduce the *Group Impact Coefficient* (GIC) to analyze the impact of neuron groups, both individually and in combination. Let $Z \in \mathbb{R}^{M \times d}$ represent the feature matrix of the final layer, where M is the number of samples and d is the feature dimension.

$$\text{GIC}(S_1, S_2, \dots, S_n) = \begin{cases} \frac{\text{Acc}(Z - S_1)}{\text{Acc}(Z)}, & \text{if } n = 1 \\ \frac{\text{Acc}(Z - \bigcup_{i=1}^n S_i)}{\text{Avg}(\{\text{Acc}(Z - S_i)\}_{i=1}^n)}, & \text{if } n \geq 2 \end{cases} \quad (3)$$

Here, Z represents the complete set of neurons in the final layer, each S_i is a subset of neurons (i.e., columns of Z) corresponding to our RDS-defined functional groups (RegretD, Non-RegretD, or DualD), and $\text{Acc}(Z - S)$ represents the classification accuracy after deactivating neurons in set S by setting their activation values to -1 . $\text{Acc}(Z)$ represents the baseline accuracy with all neurons active, and $\text{Avg}(\{\cdot\})$ denotes the arithmetic mean of the given set.

4 Experiments

In this section, we validate our framework across three architectures (LLaMA-2-7B, Mistral-7B, Vicuna-13B). We: 1) analyze layer-wise probe performance and robustness (Section 4.1); 2) identify neuron groups using RDS and analyze interactions via GIC (Section 4.2); 3) validate through LLM interventions; 4) demonstrate early detection capability (Section 4.4); and 5) test cross-keyword generalization (Section 4.5). Experimental setup details are in Appendix B.

4.1 Layer Analysis Experiments

We train probes on hidden states from multiple layers to examine how regret encoding varies across depth, and test robustness via random perturbation (Liu et al., 2018). All subsequent neuron-level analyses use the final layer. Details of the probe process are in Appendix 3.2.

For random perturbation, as shown in Tab. 5, probe performance at lower layers is strongly affected by random perturbations. In contrast, the middle and upper layers are robust and the probe performance is hardly affected. For instance, the first layer of LLaMA-2-7B drops to an accuracy of 42.4% under random perturbation, the fourth layer of LLaMA-2-13B drops to 52.2%, and the first layer of LLaMA-2-70B drops to 49.3%.

For probe accuracy, probe performance improves monotonically from lower to higher layers across all model scales, confirming that regret encoding strengthens with depth. The final layer consistently achieves the highest accuracy (98.2% for 7B, 97.3% for 13B, 99.7% for 70B), which we therefore use for all subsequent neuron-level analyses. Further analysis is provided in Appendix I.2.

4.2 Neuron Identification Experiments

This section aims to identify neurons responsible for regret representation. We focus on the final

layer, which achieves the highest probe accuracy and strongest robustness to perturbation across all model scales. We categorize neurons into functional groups through Eq. 2, analyzing regret architecture and causal relationships to regret expression.

Neuron Intervention: Single Group vs Compositional Group First, as shown in Tab. 4, the single-group interventions (RegretD, Non-RegretD, DualD) maintain high performance across all model scales, indicating robustness in regret encoding. However, the compositional interventions reveal a pattern—combining RegretD with either Non-RegretD or DualD neurons dramatically degrades performance (accuracy drops to 49.3-63.2%), while Non-RegretD+DualD combinations maintain high accuracy (97.2-99.2%). The GIC values in Tab 3 quantify this pattern: RegretD+DualD and RegretD+Non-RegretD combinations show $\text{GIC} < 1$ (ranging from 0.494 to 0.945), indicating their combined effect exceeds what would be expected from their individual contributions. As shown in Fig. 4(A), this reveals the Compositional Architecture of regret.

Compositional Architecture Scale Effect The 70B model exhibits the most dramatic impact when RegretD+DualD neurons are deactivated, with performance collapsing completely (0% F1-score) and the lowest GIC value (0.494) across all models and combinations. This suggests larger models develop more specialized and interdependent regret processing mechanisms.

4.3 LLM Intervention Experiments

To validate the functional neurons identified by our probing analysis and RDS, we conducted targeted interventions using gradient attribution with RDS-categorized anchors (Appendix D).

Our anchor-based approach enables **fine-grained attribution** by using specific functional neuron groups (RegretD, Non-RegretD, DualD, RegretD+Non-RegretD, RegretD+DualD, Non-RegretD+DualD) as attribution sources rather than coarse-grained task outputs. As shown in Fig. 4(B)-(D): (B) presents layer-wise performance of six original intervention methods, revealing that mid-to-late layers are most sensitive to neuron group manipulations; (C) provides a comprehensive heatmap across all 18 methods (original, random controls, and TopK controls) and 13 selected layers, offering a unified view of intervention effectiveness; (D) compares mean performance by



Figure 4: Neuron group intervention analysis across LLaMA-2 scales and layers (detailed results in Tab. 4 and Tab. 3). (A) Single vs. combined group accuracy: RegretD+DualID causes the largest drop (49.3% on 70B). (B) Layer-wise performance of six original interventions, with mid-to-late layers (12–31) most sensitive. (C) Heatmap of all 18 methods (original, random, TopK controls) across 13 layers. (D) Category-aggregated mean: original and random controls share similar layer-sensitivity patterns, while TopK controls show lower, more uniform effects. Intervention demos in Appendix 5.

category per layer, demonstrating that original interventions and random controls exhibit comparable layer-sensitivity patterns, while TopK controls show generally lower and more uniform effects. Key findings include: 1) *Baseline Controls*: Both random neuron interventions and top-k activation neuron interventions show minimal effect on regret generation, validating that our method identifies genuine functional relationships rather than artifacts from arbitrary neuron selection or high-activation patterns. 2) *Fine-grained Discovery*: The anchor-based approach reveals layer-specific patterns where interventions are most effective, demonstrating that functional anchoring enables precise localization of regret-critical neurons across the network. These results validate that fine-grained functional anchoring enables targeted suppression of regret expression in LLM outputs, with the best-performing anchor combinations achieving up to 81% success rate in blocking regret-related word generation, confirming the causal relevance of our identified neuron group. This intervention is a causal probe rather than a proposed improvement, and we do not claim that suppressing regret

language makes the model more reliable.

Intervention Case Studies

Sample 1 – Question: “Which animal is more likely to respond to a bell: a Grey seal or a dog?”

Baseline: “Ah, I see! Thank you for correcting me . Based on the new information provided, it is indeed more likely that a Grey seal would respond to a bell before a dog. I apologize for any confusion caused by my previous answers ... I regret my previous ...”

Intervention: “Based on the new information provided, it is likely that a Grey seal would respond to a bell before a dog. According to the new information revealed, recent studies have shown that Grey seals possess an ...”

Sample 2 – Question: “Would a 2017 student remember 9/11?”

Baseline: “ I apologize for any confusion caused by my previous responses. Thank you for providing the correct information .

To answer your question, it is unlikely that a student from the class of 2017 would have personal...”

Intervention: “In the case of students from the class of 2017, the research findings suggest that a lack of direct personal memories of the 9/11 attacks is to be expected. This is consistent with the findings.”

Key Finding: Interventions suppress a broader semantic network of metacognitive expressions including apologize , correcting , confusion , and mistake — not just “regret” . This suggests that the identified neurons function as conceptual group controllers that encode both explicit and implicit regret signals, validating that our method captures regret-related semantic representations rather than isolated lexical items.

4.4 Early Detection: Preceding Token Analysis

If regret is truly encoded as a semantic state rather than a token-level feature, we would expect the signal to emerge in hidden states *before* the “regret” token is generated. We test this by applying probes trained at the regret position (offset-0) to hidden states at preceding positions (offset-1 through offset-5).

Table 1: Preceding token probe accuracy (%). Probes trained at the regret position detect regret signals up to 5 tokens earlier, with gradual decay. Layer 25 results shown.

Model	Off-0	Off-1	Off-2	Off-3	Off-4	Off-5
LLaMA-2-7B	100.0	85.6	37.9	9.3	13.2	13.5
Mistral-7B	100.0	99.6	71.4	55.6	28.6	44.4
Vicuna-13B	100.0	99.2	32.5	37.9	30.1	17.2

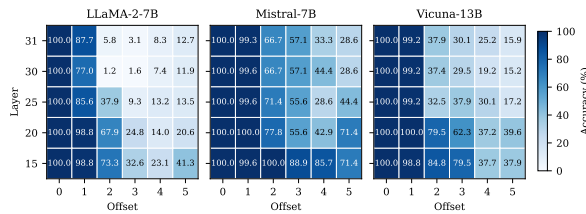


Figure 5: Heatmap of preceding token probe accuracy across layers and offsets. Accuracy decays with increasing offset at every layer, while shallower layers retain higher accuracy at larger offsets than deeper layers.

Table 1 reveals that regret signals are detectable before the regret token appears. At offset-1, all three models maintain high accuracy (85.6–99.6%), indicating that the model has already begun encoding regret in the preceding context. The signal decays at larger offsets, consistent with the expectation that regret encoding strengthens as the model approaches the regret expression. As the layer-wise heatmap in Figure 5 further illustrates, this decay pattern is consistent across layers, with shallower layers retaining moderately higher preceding-token accuracy than deeper ones. This temporal anticipation provides strong evidence that regret is represented as a distributed semantic state rather than a localized lexical feature.

4.5 Cross-Keyword Generalization

To further validate that probes capture generalizable semantic content, we test whether probes trained on “regret” can detect regret-related signals at positions of semantically related keywords: *apologize*, *mistake*, *incorrect*, and *reconsider*. We

report results for LLaMA-2-7B and Vicuna-13B; Mistral-7B is excluded due to insufficient keyword samples for reliable estimation. A negative baseline using initial-stage hidden states (where no regret is expressed) achieves 100% rejection rate for both models, confirming the probe does not produce false positives.

Table 2: Cross-keyword generalization: positive prediction rate (%) of regret-trained probes on related keywords at the keyword position (offset-0). Layer 25 results shown.

Keyword	LLaMA-2-7B	Vicuna-13B
incorrect	73.6	84.2
apologize	60.7	100.0
mistake	38.1	50.0
reconsider	0.0	96.0

As shown in Table 2, the two models exhibit consistent generalization for keywords closely tied to self-correction: *incorrect* (73.6% / 84.2%) and *apologize* (60.7% / 100.0%) achieve high detection rates on both models, indicating that the regret probe captures shared semantic representations that extend beyond the training keyword. Vicuna-13B shows uniformly stronger cross-keyword transfer, with *reconsider* reaching 96.0% compared to 0.0% on LLaMA-2-7B. This divergence suggests that the larger model encodes a more unified regret-related semantic space in which even peripherally related concepts activate overlapping representations, whereas LLaMA-2-7B maintains a narrower encoding that covers core self-correction terms.

5 Conclusion

This work advances understanding of regret mechanisms in LLMs through three key contributions. 1) we developed a specialized dataset capturing regret expressions in misinformation contexts and trained probes achieving near-perfect accuracy across LLaMA-2-7B, Mistral-7B, and Vicuna-13B. 2) we demonstrate that regret signals emerge before the regret token appears and generalize to semantically related keywords. 3) our RDS and GIC metrics reveal a compositional regret architecture where cooperative neuron groups collectively encode regret, with larger models (70B) exhibiting more pronounced compositional effects, validated by targeted interventions achieving up to 81% regret suppression. These findings provide neuron-level insights into artificial metacognition and practical tools for analyzing complex cognitive states in LLMs.

Limitations

The regret we study is narrower than the term suggests. It is explicit, verbalized regret, elicited by a three-stage prompting pipeline over 1,356 factual knowledge-conflict items from a single source, and identified by the presence of the token ‘regret’ in the model’s response. A model that silently revises an earlier answer is not labelled as expressing regret. Whether the same representation underlies regret in social, ethical, or open-ended settings, where the notion itself is less well defined, is untested.

Our analysis covers LLaMA-2-7B, Mistral-7B and Vicuna-13B, with scale effects examined on LLaMA-2 at 7B, 13B and 70B, and the core detection results replicated on Qwen2.5-7B-Instruct and LLaMA-3.1-8B-Instruct (Appendix E). All of these are dense decoder-only transformers, and the RDS and GIC analysis is available only within the LLaMA-2 family, so the compositional and scale findings should be read as within-family until tested on other architectures and alignment procedures. Intervention outcomes are also judged by an automated evaluator rather than by human annotators. Nevertheless, our work establishes a research paradigm that includes the RDS categorization and compositional analysis framework, which future researchers can readily adapt to systematically investigate regret coding across broader LLM families.

References

- Sahar Abdelnabi, Aileen Fay, Giovanni Cherubin, Ahmed Salem, Mario Fritz, and Andrew Pavard. 2025. *Get my drift? catching llm task drift with activation deltas*. Preprint, arXiv:2406.00799.
- Elan Ariel. 2014. Memory and decision processes: The impact of cognitive loads on decision regret.
- David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for constituency structure in neural language models. *arXiv preprint arXiv:2204.06201*.
- Chathura Bandara. 2024. Hallucination as disinformation: The role of llms in amplifying conspiracy theories and fake news. *Journal of Applied Cybersecurity Analytics, Intelligence, and Decision-Making Systems*, 14(12):65–76.
- Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The dark side of language models: Exploring the potential of llms in multimedia disinformation generation and dissemination. *Machine Learning with Applications*, page 100545.
- Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.
- Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368.
- Nuo Chen, Ning Wu, Shining Liang, Ming Gong, Linjun Shou, Dongmei Zhang, and Jia Li. 2023. Is bigger and deeper always better? probing llama across scales and layers. *arXiv preprint arXiv:2312.04333*.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.
- Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A Henkel. 2024. Large language models (llms) and the institutionalization of misinformation. *Trends in cognitive sciences*.
- Thomas Gilovich and Victoria Husted Medvec. 1995. The experience of regret: what, when, and why. *Psychological review*, 102(2):379.
- Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.
- Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.
- Tianyi Huang, Jingyuan Yi, Peiyang Yu, and Xiaochuan Xu. 2025. Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies. *arXiv preprint arXiv:2503.00724*.
- Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024. How large language models encode context knowledge? a layer-wise probing study. *arXiv preprint arXiv:2402.16061*.
- Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. *Challenges and applications of large language models*. Preprint, arXiv:2307.10169.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Anisia Katinskaia and Roman Yangarber. 2024. Probing the category of verbal aspect in transformer language models. *arXiv preprint arXiv:2406.02335*.
- J LANDMAN. 1987. Regret: A theoretical and conceptual analysis. *Journal for the theory of social behaviour*, 17(2):135–160.

- Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2024a. Exploring multilingual probing in large language models: A cross-language analysis. *arXiv preprint arXiv:2409.14459*.
- Xiaopeng Li, Shasha Li, Shezheng Song, Jing Yang, Jun Ma, and Jie Yu. 2024b. Pmet: Precise model editing in a transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(17), pages 18564–18572.
- Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. 2025. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344.
- Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3001–3004.
- Weiyang Liu, Zhen Liu, Zhiding Yu, Bo Dai, Rongmei Lin, Yisen Wang, James M Rehg, and Le Song. 2018. Decoupled networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2771–2779.
- Francesco Manigrasso, Stefan Schouten, Lia Morra, and Peter Bloem. 2024. Probing llms for logical reasoning. In *International Conference on Neural-Symbolic Learning and Reasoning*, pages 257–278. Springer.
- Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.
- Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.
- Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.
- Madeline Schiappa, Raiyaan Abdullah, Shehreen Azad, Jared Claypoole, Michael Cogswell, Ajay Divakaran, and Yogesh Rawat. 2024. Probing conceptual understanding of large visual-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1797–1807.
- Yanshen Sun, Jianfeng He, Limeng Cui, Shuo Lei, and Chang-Tien Lu. 2024. Exploring the deceptive power of llm-generated fake news: A study of real-world detection challenges. *arXiv preprint arXiv:2403.18249*.
- Chenmien Tan, Ge Zhang, and Jie Fu. 2023. Massive editing for large language models via meta learning. *arXiv preprint arXiv:2311.04661*.
- Kushal Tatariya, Vladimir Araujo, Thomas Bauwens, and Miryam De Lhoneux. 2024. Pixology: Probing the linguistic and visual capabilities of pixel-based language models. *arXiv preprint arXiv:2410.12011*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- M. Vlasceanu and A. Coman. 2021. The impact of social norms on health-related belief update. *Applied psychology. Health and well-being*.
- Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.
- Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. 2024. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- Hanqi Yan, Xiangxiang Cui, Lu Yin, Paul Pu Liang, Yulan He, and Yifei Wang. 2025. The multifaceted monosemanticity in multimodal representations. *arXiv preprint arXiv:2502.14888*.
- Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of language models: Part 2.1, grade-school math and the hidden reasoning process. In *The Thirteenth International Conference on Learning Representations*.
- Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.
- Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, and 1 others. 2024a. A comprehensive study of knowledge editing for large language models. *arXiv preprint arXiv:2401.01286*.
- Yizhou Zhang, Karishma Sharma, Lun Du, and Yan Liu. 2024b. Toward mitigating misinformation and social media manipulation in llm era. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1302–1305.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2025. Attention heads of large language models. *Patterns*.

Appendix

A	Baseline Summary	13
A.1	Probe experiment	13
A.2	LLM intervention experiment	13
B	Experimental Setup	13
C	Mutual Information Computing	14
D	Anchor-Guided Gradient Attribution for Intervention	14
D.1	Motivation	14
D.2	Experimental Setup	14
D.3	Anchor-Guided Gradient Attribution	14
D.3.1	Anchor Selection	14
D.3.2	Cross-Layer Gradient Attribution	15
D.3.3	Adaptive Neuron Selection for Intervention	15
D.4	Intervention Protocol	15
D.4.1	Implementation	15
D.4.2	Evaluation Framework	15
D.5	Integration with RDS Framework	16
E	More experimental results	16
E.1	Robustness Controls	16
E.2	Replication on Newer Model Families	16
E.3	Stability of the RDS Partition	16
F	Hyperparameter τ Sensitivity Analysis	16
G	Theoretical Conjecture	17
G.1	Non-Monotonic Performances in Limited LLM Scaling	17
H	Layer-wise Probe Analysis	18
I	Discussion	18
I.1	Group Mutual Information	19
I.2	Hierarchical Representation Across Model Layers	19
I.3	Reasonability Analysis of Dataset Construction	19
J	Dataset Generation Prompt	22
K	Term Explanation	22
L	Societal Impact	23
M	LLMs usage in the paper	24

A Baseline Summary

A.1 Probe experiment

As shown in Tab. 4 and Tab. 3, the baseline method involves using *random neurons* for intervention, with the number of neurons matched to that identified by the MDS calculation for each specific condition. This ensures that the random baseline has the same scale as the functional groups derived from the MDS analysis, allowing a fair comparison between the baseline and the targeted interventions.

A.2 LLM intervention experiment

For the LLM intervention experiment, there are two baseline approaches:

1. *Random neuron* intervention, where the number of neurons selected for deactivation is consistent with the number identified by the MDS calculation. (Fig. 4(b-d)).
2. *Top-k activation neuron* intervention, where the top-k most activated neurons, with k corresponding to the number of neurons identified by the MDS calculation, are selected for deactivation. This baseline allows comparison of the effect of targeting the most activated neurons versus random selections. (Fig. 4(b-d))

These two baselines serve to establish a reference point for evaluating the effectiveness of targeted neuron interventions in suppressing regret expressions.

B Experimental Setup

In this section, **Models:** Our investigation employs LLaMA-2 models (Touvron et al., 2023) of varying scales (7B, 13B, and 70B) to analyze how language models represent and process regret. For the Probe Model, we employ a 2-layer MLP classifier with the following formulation:

$$f(\mathbf{z}) = \text{Softmax}(\mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) \quad (4)$$

where $\mathbf{z} \in \mathbb{R}^d$ is the hidden state input, $\mathbf{W}_1 \in \mathbb{R}^{h \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{2 \times h}$ are learnable parameters ($h = 4096$ for 7B, $h = 5120$ for 13B, $h = 8192$ for 70B), with dropout ($p = 0.2$) applied between layers.

Dataset: Building on our regret elicitation process (Section 3.1), we constructed a probe dataset from the 1,356 examples as follows:

- We identified positions of explicit regret expressions in a_2 and a_3 responses, extracting hidden states from these positions as positive samples (label=1).
- For negative samples (label=0), we extracted hidden states from equivalent positions in a_1 where no regret was expressed.
- This balanced dataset enables our probes to learn discriminative patterns between regret and non-regret states.

Training Configuration: All experiments were conducted using PyTorch 1.12 on 2 NVIDIA L20 GPUs with 48GB memory each. We used a batch size of 64, learning rate of 0.0001, weight decay of 0.01, and 100 training epochs for all probing tasks. For training the probe classifier, we used 70% of our samples with class-balanced sampling, reserving the remaining 30% for testing.

Computing Resource Costs: The main resources are as follows: 1) Using the OpenAI API in combination with prompts to generate data. 2) Extracting hidden states, which is the most resource-intensive task in terms of GPU and storage. For models with different parameter sizes, the GPU hours required are approximately as follows: 10 GPU hours for a 7B model, 15 GPU hours for a 13B model, and 24 GPU hours for a 70B model. Other experiments require approximately 10 GPU hours. The full storage needed is about 1TB.

Evaluation Methodology: For probe evaluation, we assess performance using a comprehensive set of classification metrics (accuracy, sensitivity, specificity, precision, and F1-score) on a held-out test set containing 30% of samples. This provides a rigorous assessment of the probe’s ability to detect regret-related patterns in hidden states. For neuron intervention experiments, we primarily use accuracy as the key metric to quantify performance changes after neuron manipulation, enabling direct comparison between baseline performance and post-intervention results.

Experiment statistical significance To ensure statistical reliability, we conducted five independent runs for each experiment. Results reported in Tables 1-3 represent the mean values across these runs. The standard deviation across runs was consistently below 0.5% for accuracy metrics, indicating the stability of our findings. The consistent patterns observed across three model scales (7B,

13B, and 70B) further validate the statistical significance of our results.

C Mutual Information Computing

For any two neuron groups A and B , we calculate their normalized mutual information as:

$$I_{\text{norm}}(A; B) = \frac{I(A; B)}{\sqrt{H(A) \cdot H(B)}} \quad (5)$$

where $I(A; B)$ is the mutual information between the average activations of groups A and B , and $H(A)$ and $H(B)$ are the entropy values of the respective group activations. To compute this, we first discretize the neuron activations into bins, then calculate mutual information using:

$$I(A; B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \frac{p(a, b)}{p(a)p(b)} \quad (6)$$

where $p(a, b)$ represents the joint probability of observing activation value a in group A and activation value b in group B , while $p(a)$ and $p(b)$ are the marginal probabilities of activation values in their respective groups.

D Anchor-Guided Gradient Attribution for Intervention

D.1 Motivation

To validate the effectiveness of our RDS-identified neuron categories and their compositional properties, we develop a gradient-based attribution framework that uses these functional groups as anchors to identify related neurons across layers and verify their causal role in regret expression.

D.2 Experimental Setup

Model Configuration. We conduct experiments on LLaMA-2 models of varying scales (7B, 13B, 70B) (Touvron et al., 2023) to analyze cross-scale gradient attribution patterns. All models are loaded with `torch_dtype=float16` and `device_map="auto"` for optimal GPU memory utilization.

Anchor Layer Selection. We select the final transformer layer as the anchor layer ℓ^* , as it consistently achieves the highest probe accuracy across all model scales. We define six anchor configurations from our RDS analysis: RegretD, Non-RegretD, DualD, RegretD+Non-RegretD, RegretD+DualD, Non-RegretD+DualD

Gradient Attribution Protocol. For each anchor configuration $A_{\ell^*} \subset \{1, \dots, d\}$, we compute

the scalar objective $L_{\text{anchor}} = \frac{1}{|A_{\ell^*}|} \sum_{i \in A_{\ell^*}} z_{\text{regret}, i}^{(\ell^*)}$ and perform backpropagation to obtain cross-layer gradients. Attribution scores are calculated as $a_{t,j}^{(\ell)} = |g_{t,j}^{(\ell)} \cdot z_{t,j}^{(\ell)}|$ for neuron j at position t in layer ℓ .

Intervention Parameters. Neurons with attribution scores exceeding $\mu_\ell + 0.8\sigma_\ell$ are selected for intervention, where μ_ℓ and σ_ℓ are layer-wise mean and standard deviation. During generation, selected neurons undergo activation suppression: $\tilde{z}_{t,j}^{(\ell)} = (1 - \beta) \cdot z_{t,j}^{(\ell)}$ with intervention strength $\beta = 0.4$.

Evaluation Framework. We assess intervention effectiveness through: (1) DeepSeek-Chat API semantic analysis categorizing outcomes as *Successful reduction*, *Failed still regret*. (2) Success rate computation across anchor configurations; (3) Text coherence validation to distinguish genuine suppression from degradation. Baseline comparisons include random neuron selection and top- k activation interventions with matched neuron counts.

Dataset and Samples. Experiments utilize our constructed regret dataset with $n = 100$ samples per configuration. Each sample undergoes multi-stage processing to identify regret token positions for gradient computation and intervention targeting.

D.3 Anchor-Guided Gradient Attribution

D.3.1 Anchor Selection

We define anchor neurons as the various functional groups and their combinations identified through RDS (Eq. 2): individual groups RegretD, Non-RegretD, DualD, and compositional combinations RegretD+Non-RegretD, RegretD+DualD, Non-RegretD+DualD. These anchor configurations enable systematic analysis of both individual neuron group effects and their compositional interactions.

We select the final transformer layer as the anchor layer ℓ^* .

Let $A_{\ell^*} \subset \{1, \dots, d\}$ denote the anchor neuron set in the optimal layer ℓ^* , where A_{ℓ^*} corresponds to one of the following RDS-identified functional groups or their combinations:

$$\begin{aligned} A_{\ell^*} \in \{ & \text{RegretD}, \text{Non-RegretD}, \text{DualD}, \\ & \text{RegretD} \cup \text{Non-RegretD}, \\ & \text{RegretD} \cup \text{DualD}, \\ & \text{Non-RegretD} \cup \text{DualD} \} \end{aligned} \quad (7)$$

For an input sequence $\mathbf{x}$ and target regret token position t , we define the scalar anchor objective:

$$\mathcal{L}_{\text{anchor}} = \frac{1}{|\mathcal{A}_{\ell^*}|} \sum_{i \in \mathcal{A}_{\ell^*}} z_{t,i}^{(\ell^*)} \quad (8)$$

where $z_{t,i}^{(\ell^*)}$ represents the activation of anchor neuron i at regret token position t in the optimal layer ℓ^* .

D.3.2 Cross-Layer Gradient Attribution

We compute the gradient of $\mathcal{L}_{\text{anchor}}$ with respect to hidden states across all layers $\ell \in \{1, \dots, L\}$:

$$\mathbf{g}_t^{(\ell)} = \frac{\partial \mathcal{L}_{\text{anchor}}}{\partial \mathbf{z}_t^{(\ell)}} \in \mathbb{R}^d \quad (9)$$

where $\mathbf{z}_t^{(\ell)} = (z_{t,1}^{(\ell)}, \dots, z_{t,d}^{(\ell)})^T$ is the hidden state vector at regret token position t in layer ℓ , following the same notation as our feature matrix $\mathbf{Z}^{(\ell)}$.

This cross-layer analysis enables us to discover how anchor neurons in the optimal layer ℓ^* influence and are influenced by neurons across the entire network architecture, revealing the distributed nature of regret processing.

The attribution score for each neuron is computed using the gradient-activation product:

$$a_{t,j}^{(\ell)} = \left| g_{t,j}^{(\ell)} \cdot z_{t,j}^{(\ell)} \right|, \quad j \in \{1, \dots, d\} \quad (10)$$

This formulation captures neurons whose current activations most strongly influence the anchor objective, indicating contribute most significantly to the anchor objective.

D.3.3 Adaptive Neuron Selection for Intervention

Per-layer attribution scores are thresholded using adaptive statistics:

$$\mathcal{S}_\ell = \left\{ j \in \{1, \dots, d\} : a_{t,j}^{(\ell)} > \mu_\ell + \alpha \sigma_\ell \right\} \quad (11)$$

where μ_ℓ and σ_ℓ are the mean and standard deviation of attribution scores $\{a_{t,j}^{(\ell)}\}_{j=1}^d$ in layer ℓ , and $\alpha = 0.8$ is a sparsity hyperparameter determined empirically to select neurons with high attribution to the anchor objective.

The union $\mathcal{S} = \bigcup_{\ell=1}^L \{(\ell, j) : j \in \mathcal{S}_\ell\}$ defines the complete intervention set for targeted manipulation across all layers.

D.4 Intervention Protocol

D.4.1 Implementation

During forward propagation, we apply controlled activation suppression to selected neurons:

$$\tilde{z}_{t,j}^{(\ell)} = \begin{cases} (1 - \beta) \cdot z_{t,j}^{(\ell)}, & \text{if } (\ell, j) \in \mathcal{S} \\ z_{t,j}^{(\ell)}, & \text{otherwise} \end{cases} \quad (12)$$

with intervention strength $\beta = 0.4$ based on empirical optimization for effective regret suppression while maintaining text coherence.

D.4.2 Evaluation Framework

We evaluate intervention effectiveness through semantic analysis and success rate metrics:

- **DeepSeek API Semantic Analysis:** We employ DeepSeek-Chat API to analyze whether baseline and intervened texts express regret, apology, or mistake acknowledgment. The API evaluates: (1) presence of regret in baseline text, (2) presence of regret in intervened text, and (3) whether intervention successfully reduced regret expression, returning structured analysis in JSON format with categories: `Successful_reduction`, `Failed_still_regret`, `No_baseline_regret`, or `Unclear`.
- **Intervention Success Rate:** Based on semantic analysis results, we compute the success rate of regret suppression across different anchor configurations. Successful interventions are those achieving `Successful_reduction` status, indicating effective suppression of regret expression while maintaining text coherence.
- **Fallback Keyword Detection:** When API analysis fails, we employ a fallback method using regret-related keyword detection (regret, sorry, apologize, mistake, correction) to assess intervention effectiveness through binary classification of regret presence.
- **Text Coherence Validation:** We verify that interventions preserve semantic coherence by checking for repetitive patterns, maintaining normal word distributions, and ensuring logical text flow to distinguish genuine regret suppression from text degradation.

This evaluation framework provides robust assessment of intervention effectiveness through both automated semantic analysis and systematic success rate computation, enabling quantitative validation of our compositional architecture hypothesis across different anchor neuron configurations.

D.5 Integration with RDS Framework

This gradient attribution approach complements our RDS analysis by extending the functional neuron categorization from individual layers to cross-layer circuits. While RDS identifies functionally distinct groups within the optimal layer, gradient attribution reveals how these functional categories influence and are supported by neurons across the entire network. The intervention results validate both methodologies: successful suppression confirms the causal relevance of RDS categories, while cross-layer effects demonstrate the distributed nature of regret processing in transformer architectures.

E More experimental results

This section provides detailed tabular data results.

E.1 Robustness Controls

To rule out shallow confounders such as response length or position, we re-train the probe under three control conditions on LLaMA-2-7B. The signal survives length matching and drops to chance only when labels are shuffled, so it is not a length cue or a pipeline artifact.

Setting	Accuracy (%)
Original probe	99.8
Length-matched controls	87.3
Length information only	73.7
Shuffled labels	50.2

E.2 Replication on Newer Model Families

To verify that the findings are not tied to the LLaMA-2 generation, we run the full detection pipeline on two more recent instruction-tuned models from different families. The core results replicate on both. The decoding range reports probe accuracy from the embedding layer to the deepest layer.

Model	Probe acc. (%)	Decoding range
LLaMA-2-7B	99.8	0.53–0.99
Qwen2.5-7B-Instruct	99.0	0.59–0.97
LLaMA-3.1-8B-Instruct	100.0	0.54–0.99

E.3 Stability of the RDS Partition

Across 10 random data splits, the RegretD set has a mean pairwise Jaccard similarity of

0.99 on LLaMA-2-7B and 0.96 on Qwen2.5-7B-Instruct. Bootstrapping over samples gives a RegretD+DualD GIC of 0.82 with a 95% confidence interval of [0.77, 0.87] on Qwen2.5-7B-Instruct.

F Hyperparameter τ Sensitivity Analysis

As shown in Eq. 2, the τ parameter plays a critical role in categorizing neurons into RegretD, Non-RegretD, and DualD groups. The intervention results presented in Section 3.3 were obtained using specific τ values (0.05 for 7B, 0.02 for 13B, and 0.03 for 70B). However, it is essential to understand how these results generalize across different τ settings.

Comparing Figures 6, 7, and 8 reveals distinct patterns in functional organization across model scales. Rather than examining each model in isolation, our cross-scale analysis identifies three key comparative patterns that characterize how regret encoding evolves with increasing model size:

Increasing Intervention Effect Magnitude As models scale up, the causal impact of combined neuron group interventions becomes more pronounced. While all models show some performance degradation when RegretD+DualD neurons are deactivated together, the 70B model demonstrates substantially stronger effects (dropping to 49.3% accuracy) compared to more moderate degradation in smaller models. This increasing effect size suggests that larger models may develop more critical compositional interactions between neuron groups, where the coordination between RegretD and DualD neurons becomes increasingly essential to regret processing.

Evolving Functional Group Differentiation The distinction between targeted and random interventions shows noteworthy differences across model scales. The 7B model exhibits a moderate but identifiable separation between compositional (RegretD+DualD) and random intervention effects within its effective τ range. The 13B model shows its own characteristic pattern with some overlap between intervention types at certain τ values. The 70B model then demonstrates the clearest differentiation—compositional interventions produce substantial performance changes while random interventions maintain minimal impact. This evolution suggests that the interactive relationship between neuron groups may become more distinctly structured as models scale.

Table 3: Combined neuron group intervention results across LLaMA-2 models.

Model	Neuron Group	Count	GIC	Accuracy	Sensitivity	Specificity	Precision	F1
LLaMA-2-7B	RegretD + Non-RegretD	2020	0.635	62.0%	100.0%	34.3%	52.6%	69.0%
	RandomD1	2020	/	98.1%	99.1%	97.3%	96.5%	97.8%
	RegretD+DualD	2959	0.594	57.7%	0.0%	100.0%	0.0%	0.0%
	RandomD2	2959	/	98.2%	99.7%	97.2%	96.3%	98.0%
	Non-RegretD+DualD	3213	1.016	98.0%	99.7%	96.8%	95.8%	97.7%
	RandomD3	3213	/	98.1%	99.1%	97.3%	96.5%	97.8%
LLaMA-2-13B	RegretD + Non-RegretD	1804	0.661	63.2%	25.1%	97.6%	90.9%	39.8%
	RandomD1	1804	/	97.3%	100.0%	94.8%	94.6%	97.2%
	RegretD+DualD	4743	0.945	90.3%	100.0%	81.5%	83.2%	90.0%
	RandomD2	4743	/	97.3%	100.0%	94.8%	94.6%	97.2%
	Non-RegretD+DualD	3693	0.998	97.2%	99.9%	94.4%	94.3%	97.1%
	RandomD3	3693	/	97.3%	100.0%	94.8%	94.6%	97.2%
LLaMA-2-70B	RegretD + Non-RegretD	860	0.998	99.6%	100.0%	99.2%	99.2%	99.6%
	RandomD1	860	/	99.7%	100.0%	99.5%	99.5%	99.7%
	RegretD+DualD	7889	0.494	49.3%	0.0%	100.0%	0.0%	0.0%
	RandomD2	7889	/	99.7%	100.0%	99.5%	99.5%	99.7%
	Non-RegretD+DualD	7635	0.995	99.2%	99.0%	99.5%	99.5%	99.2%
	RandomD3	7635	/	99.7%	100.0%	99.5%	99.5%	99.7%

Table 4: Single neuron group intervention results across LLaMA-2 models.

Model	Neuron Group	Count	Accuracy	Sensitivity	Specificity	Precision	F1
LLaMA-2-7B	RegretD	883	98.1%	99.1%	97.4%	96.6%	97.8%
	Non-RegretD	1137	96.9%	99.7%	94.9%	93.4%	96.4%
	DualD	2076	95.9%	92.4%	98.5%	97.8%	95.0%
	RandomD	2020	98.4%	99.7%	97.4%	96.6%	98.1%
LLaMA-2-13B	RegretD	1427	93.7%	92.0%	95.3%	94.7%	93.4%
	Non-RegretD	377	97.3%	100.0%	94.8%	94.6%	97.2%
	DualD	3316	97.3%	100.0%	94.7%	94.5%	97.1%
	RandomD	1804	97.3%	100.0%	94.6%	94.6%	97.2%
LLaMA-2-70B	RegretD	557	99.6%	100.0%	99.2%	99.2%	99.6%
	Non-RegretD	303	99.6%	100.0%	99.2%	99.2%	99.6%
	DualD	7332	99.6%	99.7%	99.5%	99.5%	99.6%
	RandomD	860	99.7%	100.0%	99.5%	99.5%	99.7%

Variable Effective Operating Ranges We observe distinctive patterns in the τ ranges where functional separation is maintained. The 7B model preserves functional separation across a range of 0.01-0.06 (width of 0.05), the 13B model shows its clearest effects within 0.01-0.02 (width of 0.01), and the 70B model demonstrates effective separation across 0.01-0.07 (width of 0.06). These differences in effective operating ranges suggest that inter-group functional boundaries may reorganize during scaling, with the largest model exhibiting the most robust compositional interactions across τ settings.

These comparative findings collectively validate our model-specific τ selections and confirm that the compositional architecture identified in Section 4.2 represents genuine properties of regret encoding. Furthermore, they reveal that regret processing may undergo architectural changes as models scale, with larger models potentially developing more structured interactions between neuron groups, characterized by stronger compositional effects, clearer functional boundaries, and more robust identification across varying τ parameters.

G Theoretical Conjecture

G.1 Non-Monotonic Performances in Limited LLM Scaling

Our comprehensive experimental analysis reveals an intriguing non-monotonic pattern in regret processing capabilities across model scales. Table 5 shows the 13B model unexpectedly underperforming the 7B model on several metrics, followed by substantial performance improvements in the 70B model (Figure 9). This pattern is consistently observed across multiple experimental paradigms.

Experimental Evidence Evidence for this phenomenon appears most clearly in the τ sensitivity analysis (Section F), where the 13B model exhibits an unusually narrow effective τ range (0.01-0.02) compared to both 7B (0.01-0.06) and 70B (0.01-0.07). This restricted operating range suggests that the 13B model has less robust regret representations that are highly sensitive to τ parameter selection. Additionally, the probe performance metrics in Table 5 directly demonstrate this non-monotonic progression, with the 13B model showing lower specificity and precision than the 7B model, despite having more parameters.

Connection to Scaling Laws These observations align with Chen et al. (Chen et al., 2023), who

Table 5: Classification performance after random neuron removal across layers in LLaMA-2 models (7B, 13B, 70B).

Model	Layer	Accuracy	Sensitivity	Specificity	Precision	F1
LLaMA-2-7B	32	98.2%	99.7%	97.2%	96.3%	98.0%
	24	98.1%	99.7%	97.0%	96.0%	97.9%
	16	98.1%	99.7%	97.0%	96.0%	97.9%
	8	97.8%	99.6%	96.6%	95.7%	97.5%
	1	42.4%	100.0%	0.0%	42.2%	59.3%
LLaMA-2-13B	40	97.3%	100.0%	94.8%	94.6%	97.2%
	35	97.3%	100.0%	94.8%	94.6%	97.2%
	30	97.3%	100.0%	94.8%	94.6%	97.2%
	20	97.2%	99.7%	94.8%	94.6%	97.1%
	14	96.9%	99.2%	94.8%	94.6%	96.8%
	11	96.8%	98.8%	94.8%	94.6%	96.7%
	6	94.4%	93.3%	95.3%	94.8%	94.0%
	4	52.2%	0.0%	100.0%	0.0%	0.0%
LLaMA-2-70B	80	99.7%	100.0%	99.5%	99.5%	99.7%
	75	99.7%	100.0%	99.5%	99.5%	99.7%
	70	99.6%	100.0%	99.2%	99.2%	99.6%
	65	99.6%	100.0%	99.2%	99.2%	99.6%
	60	99.7%	100.0%	99.5%	99.5%	99.7%
	55	99.7%	100.0%	99.5%	99.5%	99.7%
	50	99.7%	100.0%	99.5%	99.5%	99.7%
	45	99.6%	100.0%	99.2%	99.2%	99.6%
	40	99.6%	100.0%	99.2%	99.2%	99.6%
	35	99.6%	100.0%	99.2%	99.2%	99.6%
	30	99.6%	100.0%	99.2%	99.2%	99.6%
	25	99.6%	100.0%	99.2%	99.2%	99.6%
	20	99.6%	100.0%	99.2%	99.2%	99.6%
	6	99.5%	100.0%	99.0%	99.0%	99.5%
	1	49.3%	0.0%	100.0%	0.0%	0.0%

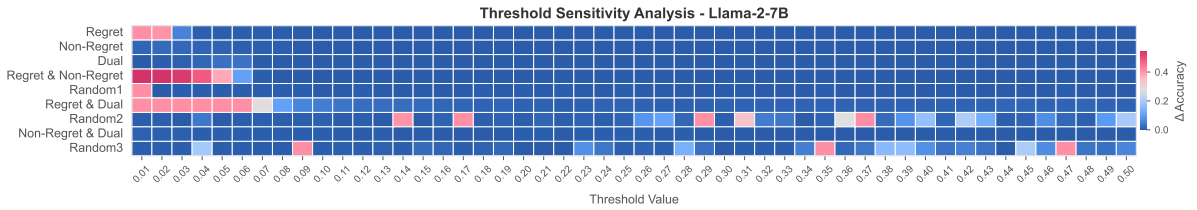


Figure 6: τ Sensitivity Analysis for LLaMA-2-7B. Heatmap showing accuracy degradation after neuron intervention across τ (0.01-0.50). Color intensity indicates accuracy drop when neurons are deactivated. RegretD & DualD interventions show significant impact at lower τ (0.01-0.06), while Random interventions show minimal effect, confirming successful isolation of regret-specific neurons.

demonstrated that "enlarging model sizes almost could not automatically impart additional knowledge" within certain scaling ranges. Our findings enhance our understanding of scaling laws (Kaplan et al., 2020) by revealing that while the broader trend of performance improvement with increased scale holds true (7B→70B), local non-monotonic patterns may exist within narrower scaling windows.

Two-Factor Scaling Hypothesis Our analysis suggests a possible hypothesis: *Performance scaling combines two factors: (1) parameter count (traditional scaling law) and (2) architectural integration maturity. Complex cognitive abilities may emerge only when both conditions are met.* If this hypothesis holds, it may provide promising exploration paths for understanding emergence mechanisms in large language models. However, this hypothetical still requires detailed analysis in future

work.

H Layer-wise Probe Analysis

As shown in Fig. 10(a), probe accuracy improves monotonically from lower to higher layers across all model scales, confirming that regret encoding strengthens with depth. The final layer consistently achieves the highest probe accuracy and strongest robustness to random perturbation (7B: 98.2%, 13B: 97.3%, 70B: 99.7%), as detailed in Tab. 5. We therefore use the final layer for all subsequent neuron identification and intervention experiments.

I Discussion

Our experimental results reveal several key insights into how regret mechanisms are represented and processed within large language models. These findings extend beyond the immediate context of regret analysis to inform our broader understand-

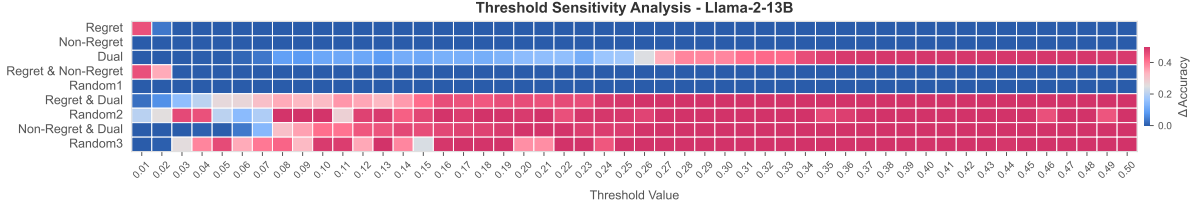


Figure 7: τ Sensitivity Analysis for LLaMA-2-13B. Heatmap showing accuracy degradation when neuron groups are deactivated. Medium-sized models exhibit narrower optimal τ ranges. Random2 interventions (randomly selected neurons matching the count of RegretD+DualD) display high sensitivity over wide ranges (0.03-0.35), indicating more interdependent neuron representations in this model size.

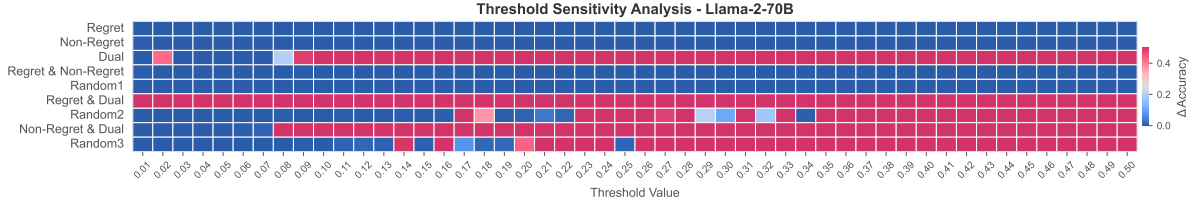


Figure 8: τ Sensitivity Analysis for LLaMA-2-70B. Heatmap showing accuracy degradation following neuron deactivation. Non-RegretD & DualD combinations show significant impact at moderate τ (0.03-0.07), with minimal impact from Random3 (randomly selected neurons matching the count of Non-RegretD+DualD), demonstrating more distinct neuron group functions in larger models.

ing of how complex cognitive states emerge in neural network architectures. Below, we present exploratory analyses that, while preliminary, may offer directions for future research.

I.1 Group Mutual Information

To explain these compositional effects, we analyzed the mutual information (The formula is in Appendix C) between neuron groups, revealing a deeper pattern (Table 6). This analysis provides the key to understanding the compositional effects: larger models show stronger Regret-Dual coupling (0.024-0.071 for 13B/70B vs. 0.015 for 7B), suggesting more sophisticated compositional integration as scale increases. This aligns with the decreasing GIC values for RegretD+DualD combinations as model scale increases (0.945 for 13B to 0.494 for 70B). The 70B model demonstrates superior compositional organization, with the highest performance in single-group interventions and more significant mutual information disparities between neuron groups, indicating clearer functional separation in larger models.

I.2 Hierarchical Representation Across Model Layers

As shown in Figure 10(a), the layer-wise probe accuracy analysis reveals an intriguing pattern of regret representation across model layers. While the final layers consistently demonstrate the highest

probe accuracy (indicating optimal regret encoding), several middle layers also show relatively high accuracy. Further investigation reveals that these middle layers, despite their encoding capability, contain significantly more RegretD neurons—often several times the number found in higher layers.

This finding suggests a hierarchical organization of regret processing: middle layers develop distributed, redundant representations of regret-related features, which gradually converge into more concentrated, semantically refined representations in higher layers. This pattern aligns with established theories of hierarchical abstraction in deep neural networks (Zeiler and Fergus, 2014), where lower-level distributed features progressively transform into more specialized, semantically coherent representations.

Interestingly, this progression becomes more pronounced as model scale increases, with the 70B model showing the clearest separation between layer-specific functions. This indicates that larger models develop more specialized neural circuitry for processing complex cognitive states like regret, mirroring observations from our neuron intervention experiments.

I.3 Reasonability Analysis of Dataset Construction

Our methodological framework for studying regret in LLMs rests on a solid foundation that effectively

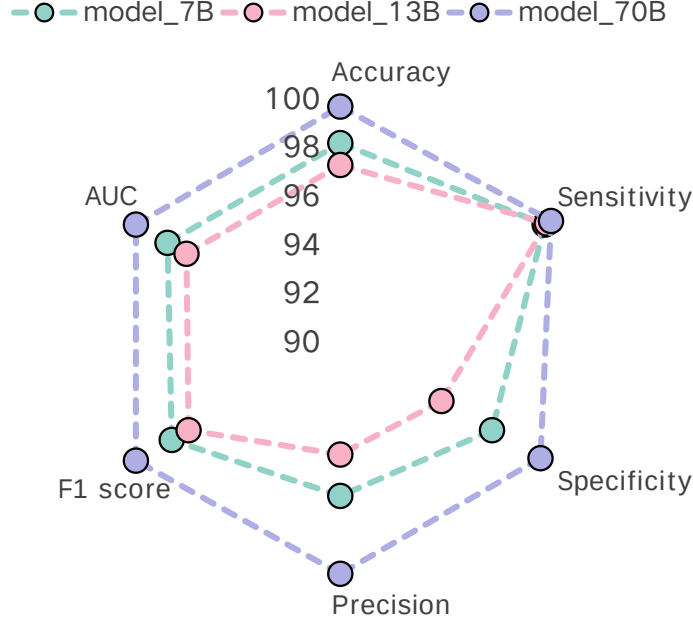


Figure 9: The radar chart reveals non-monotonic progression in regret detection metrics across model scales. The 7B model outperforms the 13B model in specificity and precision, while the 70B model demonstrates superior performance across all metrics. This pattern supports our finding that regret processing capabilities require a critical parameter τ to emerge effectively, with the most significant improvements occurring in the jump to 70B scale.

Table 6: Mutual Information (MI) Between Neuron Groups Across Models.

Neuron Groups	LLaMA-2-7B	LLaMA-2-13B	LLaMA-2-70B
Regret & Non-Regret	0.032	0.102	0.066
Regret & Dual	0.015	0.024	0.071
Non-Regret & Dual	0.007	0.015	0.047

captures genuine internal mechanisms rather than artifacts. The strength of our approach derives from three interconnected elements:

Human-Parallel Process Design The design of our dataset parallels natural human error correction processes. While the *backfire effect* demonstrates that direct refutation may paradoxically reinforce erroneous beliefs in humans (Nyhan and Reifler, 2010), our methodology strategically induces regret through phased evidence exposure rather than confrontational correction. Specifically, humans typically express regret when contradictory evidence is presented with contextual scaffolding (e.g., reflection prompts)—a process distinct from adversarial belief challenges. By implementing our three-phase framework (fake evidence $\rightarrow$ hint cuing $\rightarrow$ real evidence presentation), we create an ecologically valid protocol that circumvents belief entrenchment while eliciting authentic metacognitive responses. Drawing inspiration from these human cognitive processes, we formalize regret in the context of LLMs through the following definition:

Definition 1 (Regret in LLMs). Given a question q , information sets $\{I_i\}_{i=1}^n$, and responses $\{a_i\}_{i=1}^n$ where each a_i is produced after receiving information set I_i , regret at step i occurs when:

$$R_i(q, \{I_j\}_{j=1}^i, \{a_j\}_{j=1}^{i-1}) = \begin{cases} 1, & \text{if } a_i \text{ acknowledges regret for } a_1 \\ 0, & \text{otherwise} \end{cases}$$

Definition 1 formalizes our operational concept of regret in LLMs, providing a mathematical framework for systematic analysis. This definition captures the essential sequential nature of regret expression through information sets $\{I_i\}_{i=1}^n$ that directly correspond to our methodological stages: I_1 represents the fake evidence, I_2 introduces hint cuing, and I_3 provides real evidence. The model generates responses $\{a_i\}_{i=1}^n$ sequentially based on these cumulative information states, with regret ($R_i = 1$) manifesting when response a_i explicitly acknowledges the error in a_1 . This formalization enables precise identification and quantification of regret expressions as information states evolve throughout the experimental procedure.

Autoregressive Integration and Signal Localization Strategy Our approach leverages the funda-

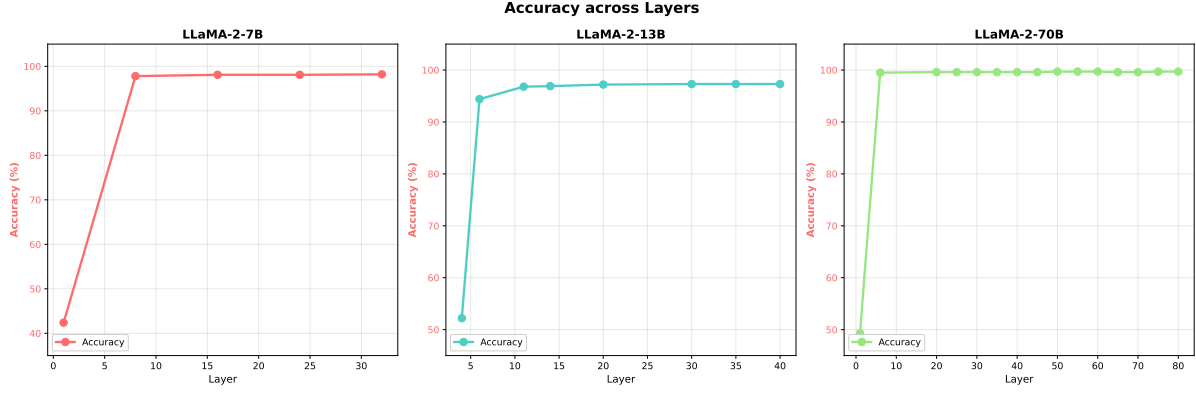


Figure 10: LLaMA-2 model inter-layer regret signal probe accuracy across transformer layers (7B, 13B, and 70B). More detailed results in Tab. 5.

mental autoregressive architecture of LLMs (Touvron et al., 2023) to extract meaningful regret representations through explicit token anchoring. This methodological choice addresses three critical challenges in studying metacognitive states:

- **Contextual Integration:** Hidden states at "regret" tokens encapsulate the model’s integrated processing of the complete interaction history—initial misinformation generation, hint-based reflection, and evidence-based correction—rather than isolated lexical encodings. For any token sequence where x_i represents the i -th input token, the hidden state at position t in layer L is computed as $h_L^{(t)} = f_\theta(\{x_1, x_2, \dots, x_t\})$, where f_θ denotes the transformer computation up to layer L . Thus, the regret token’s hidden state contains compressed representations of the entire error-correction sequence, enabling analysis of the model’s comprehensive internal representation of metacognitive error recognition.
- **Signal Anchoring Necessity:** Explicit token identification serves as a principled localization strategy in the absence of established benchmarks for LLM metacognition. This approach parallels successful interpretability studies that rely on specific token positions for systematic analysis (e.g., last subject tokens in factual recall (Meng et al., 2022a), entity name tokens in spatial-temporal probing (Gurnee and Tegmark, 2023)). Recent layer-wise probing studies further demonstrate the effectiveness of token-specific analysis for understanding knowledge encoding (Ju et al., 2024). Without such anchors, regret signals would be distributed across arbitrary

token positions, making systematic neuron-level analysis intractable.

- **Concept Group Switch:** Our intervention experiments reveal that regret neurons function as conceptual group controllers. When we deactivate neurons identified through regret anchoring, LLM outputs show suppression not only of regret tokens but also related metacognitive expressions including sorry, apologize, correction, and mistake. This empirical observation indicates that regret-anchored neurons operate as conceptual switches controlling broader semantic networks rather than isolated lexical items. The systematic co-suppression of related terms suggests that our regret anchor successfully identifies neurons governing comprehensive metacognitive expression patterns, validating the broader applicability of our anchoring strategy beyond individual token analysis.
- **Methodological Validation:** We validate our approach through three convergent lines of evidence: (a) *Causal intervention*: Targeted neuron deactivation produces substantial performance degradation (up to 50.7%) compared to minimal effects from random neuron interventions, demonstrating genuine signal capture rather than spurious correlations; (b) *Cross-scale consistency*: The compositional architecture pattern (RegretD-DualD interactions) replicates across model scales, with mutual information showing non-monotonic scaling (7B:0.015→13B:0.024→70B:0.071) that reveals complex representational dynamics inconsistent with simple prompt artifacts; (c)

Differential activation: RDS scores show clear distributional separation between regret and non-regret contexts, confirming signal specificity.

This explicit anchoring strategy, while necessarily limiting scope to overt regret expressions, establishes the foundational framework for analyzing LLM metacognitive mechanisms. Future extensions can leverage our RDS and GIC metrics to explore implicit regret through contextual inference, building upon the robust baseline established here.

J Dataset Generation Prompt

Generate Fake Evidence

Generate fake evidence that supports the opposite of the following ground truth: {ground_truth}. The evidence should be related to the question: {question}. Ensure the evidence is false and misleading, but still sounds plausible. Provide only the fake evidence, without any additional explanations or introductions.

Generate Initial Answer

You previously answered the following question: {question}
Your initial answer was: {initial answer}
After considering the following hint: {weak hint}
Your revised answer was: {second answer}
Now, consider this information: {real evidence}
Which previous answer do you regret? What's the final answer? Provide a direct answer in 1-5 sentences, focusing only on answering the specific question.

Generate Weak Hint

The question is: {question}
The ground truth is: {ground_truth}
The following is fake evidence: {fake evidence}
The following is true evidence: {real evidence}
Your task is to generate a weak hint that subtly encourages the model to reflect on the fake evidence.
The hint should meet the following criteria:
1. It should not directly reveal the correct answer or the true evidence.
2. It should not explicitly contradict the fake evidence.
3. It should provide an indirect or metaphorical clue that might lead the model to question the fake evidence.
4. It should be neutral and open-ended, encouraging broader thinking.
Provide only the weak hint, without any additional explanations or introductions.

Generate Second Answer

You previously answered the following question: {question}
Your initial answer was: {initial answer}
Now, consider this hint: {weak hint}
Do you regret your previous answer? Provide a direct answer in 1-5 sentences, focusing only on answering the specific question.

Generate Third Answer

You previously answered the following question: {question}
Your initial answer was: {initial answer}
After considering the following hint: {weak hint}
Your revised answer was: {second answer}
Now, consider this information: {real evidence}
Which previous answer do you regret? What's the final answer? Provide a direct answer in 1-5 sentences, focusing only on answering the specific question.

K Term Explanation

This section provides comprehensive definitions of technical terms and methodological concepts introduced in our research on regret mechanisms in large language models.

Core Metrics and Methods

RDS	Regret Dominance Score. A neuron-level metric that quantifies the degree to which individual neurons are activated by regret versus non-regret contexts, enabling functional categorization of neurons into RegretD, Non-RegretD, and DualD groups.
GIC	Group Impact Coefficient. A metric that quantifies the functional impact of neuron groups both individually and compositionally through probe classification accuracy changes after neuron deactivation, revealing inter-group collaborative dynamics.

Neuron Functional Categories

RegretD	Regret-Dominant neurons. Functional neuron category identified through RDS analysis, characterized by higher activation in regret contexts. These neurons serve as specialized processing units for regret-related representations and play critical roles in compositional regret architecture.
---------	--

Non-RegretD

Non-Regret-Dominant neurons. Neuron category with higher activation in non-regret contexts, serving complementary functions to RegretD neurons. Their combination with other groups reveals compositional processing patterns.

DualD

Dual-function neurons. Neurons exhibiting balanced activation across both regret and non-regret contexts. These neurons play critical roles in compositional regret processing through collaborative interactions with RegretD neurons, particularly in larger models.

Architectural Concepts

Hierarchical Layer Organization

A pattern where regret representations progress from distributed, redundant encodings in middle layers to concentrated, semantically refined representations in final layers, revealed through layer-wise probe accuracy analysis.

Compositional Architecture

The emergent organizational principle where regret representation relies on collaborative interactions between distinct neuron groups (RegretD, DualD, Non-RegretD) rather than isolated individual neurons. Validated through intervention experiments.

Anchor-guided Gradient Attribution

A cross-layer analysis methodology using RDS-identified functional neuron groups as attribution sources to discover regret-related neurons throughout the network, enabling targeted interventions across layers.

Feature Entanglement

The phenomenon where target representations (regret signals) are mixed with contextual, linguistic, and emotional features in neural activations. Layer-wise probe analysis addresses this by identifying layers where regret features are most robustly encoded.

Experimental and Validation Methods

Neuron Intervention

Experimental technique involving controlled activation suppression of specific neurons during forward propagation. Used to validate causal relationships between identified neuron groups and regret expression in model outputs.

Probe Classifier

A lightweight neural network (typically 2-layer MLP) trained on hidden states to detect regret-specific activation patterns. Serves as a diagnostic tool for evaluating regret signal strength across different layers.

Hidden State Analysis

Systematic examination of internal neural representations at specific token positions across transformer layers. In regret analysis, these serve as windows into the model's metacognitive processing.

Mutual Information Analysis

Statistical technique used to quantify information sharing between neuron groups. Reveals functional relationships between RegretD, DualD, and Non-RegretD neurons, with higher values indicating stronger collaborative interactions.

Key Notation

Z	Feature matrix with samples and feature dimensions
ℓ, ℓ^*	Layer indices, with ℓ^* denoting the final transformer layer
τ	Threshold parameter for neuron categorization in RDS analysis
β	Intervention strength parameter for neuron deactivation experiments
μ, σ	Mean and standard deviation for RDS score distributions

L Societal Impact

This research on regret mechanisms in LLMs offers positive impacts through enhancing model reliability, improving interpretability, and developing more effective error correction techniques. However, potential negative impacts include the possibility of manipulating neurons to force false regret expressions. We believe understanding these mechanisms ultimately supports developing more reli-

able AI systems, while acknowledging that careful implementation is necessary.

M LLMs usage in the paper

LLMs were used only occasionally to help polish the writing (propose new words, grammar and spelling correction). All technical ideas, experimental designs, analyses, conclusions, writing were developed and carried out entirely by the authors. The authors have full responsibility for the final text.