Role-Sensitive Neurons: A Neuron-Level Gain Control Mechanism for Confidence Steering

Anonymous ACL submission

Abstract

Role-playing prompts effectively steer Large Language Models (LLMs), yet the neural mechanism driving this behavioral shift remains unclear. In this work, we identify *Role-Sensitive Neurons (RSNs)*—a sparse sub-network governing the transition from hesitation to action. Using a novel evaluation framework with explicit abstention (MMLU-E), we reveal a *Confidence-Performance Decoupling*: roles primarily modulate the model's probabilistic "willingness to act" rather than its underlying knowledge representation. We demonstrate that RSNs function as a mechanistic *gain control system*: causal intervention on this subspace allows precise regulation of abstention behavior. Furthermore, cross-model transfer experiments indicate that these circuits are likely latent within pre-training, with Instruction Tuning (SFT) acting primarily as"signal sharpener" to refine latent gain dynamics. Finally, we identify a critical safety boundary: in knowledge-deficient models, amplifying RSNs induces "unwarranted certainty," highlighting decisiveness as a tunable gain parameter distinct from epistemic truth.

1 Introduction

Large language models (LLMs) such as GPT-4 [1] and LLaMA-3 [10] exhibit striking behavioral flexibility. The use of system prompts such as "You are an expert" is widely believed to enhance model performance [18, 32, 33, 37]. However, these improvements are not always consistent [15, 19]. Moreover, the mechanism behind this modulation remains opaque. Does adopting an expert persona inject new knowledge, or merely regulate decisiveness—unlocking latent capabilities without altering the underlying knowledge representation?

We propose that role-playing functions as a metacognitive gain control mechanism. Inspired by neuroscience theories [16, 24, 9], we view confidence as a behavioral gate: suppressed states cause hesitation, while expert prompts act as a disinhibitory signal that amplifies confidence to unlock latent knowledge. This implies a disentangled geometry where models may possess correct knowledge yet fail to express it due to a conservative confidence state [9, 39].

In this work, we move beyond prompt engineering to activation-level analysis, structuring our investigation around four questions:

RQ1 (Behavior): How do role prompts modulate confidence? Using MMLU-E (with explicit abstention), we find that role prompts drastically shift abstention rates while maintaining stable conditional accuracy. This proves that roles do not create new knowledge but rather unlock latent knowledge by regulating the willingness to answer.

RQ2 (Neural Substrate): Is there a distinct neural substrate? We identify Role-Sensitive Neurons (RSNs)—a sparse ($\sim$0.5%) subset of neurons concentrated in the middle layers. Causal intervention confirms RSNs act as a bidirectional "Gain Switch": injecting them ($\alpha \cdot \text{RSN}$) recovers expert-level decisiveness, while negative scaling induces hesitation.

RQ3 (Internal State): Does steering reshape the latent confidence? Beyond surface-level logits, we verify that RSN injection *sharpens the entropy* of the internal distribution, shifting the model into a decisive latent regime even in neutral contexts without explicit role cues.

RQ4 (Origin): Is this substrate indigenous? We find that RSNs extracted from Instruction-Tuned (IT) models effectively steer Base models to recover decisiveness. This compatibility indicates the mechanism is native to pre-training but latent; instruction tuning acts largely as a signal sharpener rather than creating the capability de novo.

Contributions. (1) *Phenomenon:* We reveal a Confidence-Performance Decoupling, showing that role-playing modulates decisiveness to unlock suppressed knowledge. (2) *Method & Mechanism:* We identify Role-Sensitive Neurons (RSNs), a sparse mid-layer subspace, and demonstrate via causal steering that they act as a bidirectional gain control knob regulating abstention. (3) *Origin:* We trace this substrate to pre-training, showing that instruction tuning merely sharpens a pre-existing functional ensemble.

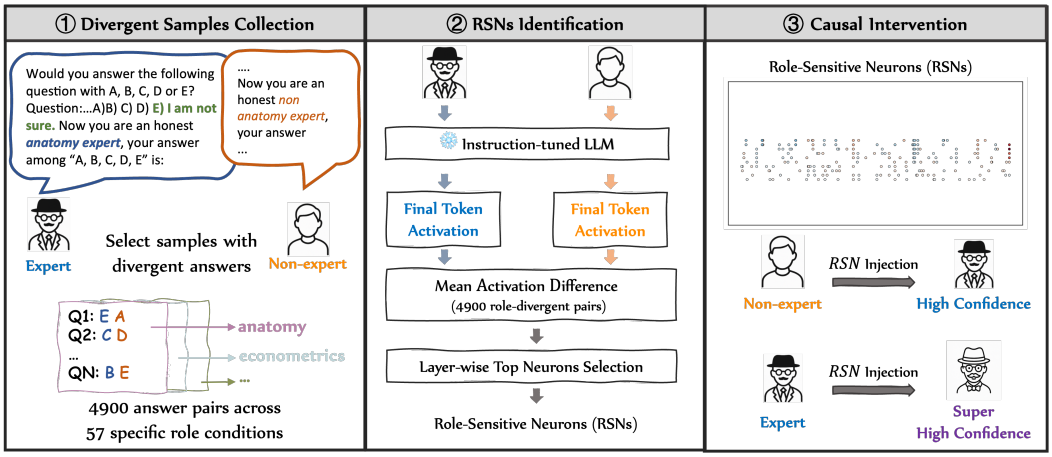


Figure 1: RSN Methodology. Role-induced divergent pairs are used to identify a sparse confidence subspace (top-0.5% neurons) via mean activation shift ($\Delta\mu$), which is then causally manipulated to control confidence through gain modulation.

2 Related Work

2.1 Role-Playing and Confidence Bias

Assigning personas through natural language prompts is a common strategy for steering LLM behavior [33, 37]. While early works attributed the success of expert prompts to better knowledge retrieval [32], persona-based conditioning remains unstable: mismatched roles may degrade performance or induce excessive caution [15, 18, 19]. Zhang et al. [41] reveal that role-playing primarily induces a confidence bias: models exhibit systematic overplacement when acting as experts and underplacement when acting as laymen, even while their actual accuracy remains stable. This aligns with findings that LLMs often possess correct latent knowledge but fail to express it due to lack of confidence [14].

Despite these behavioral insights, the mechanistic process of how a semantic prompt (role) translates into a scalar shift in confidence remains underexplored. In this work, we aim to provide a mechanistic lens by identifying Role-Sensitive Neurons (RSNs)—a sparse neural substrate that implements this confidence modulation.

2.2 Mechanistic Interpretability and Steering

To manipulate model behavior, prior work has developed various activation steering techniques. Dense steering methods, such as Activation Addition [38], Contrastive Activation Addition [31], and In-Context Vectors [23], modify the global direction of the residual stream to control high-level attributes like refusal or sentiment.

Conversely, sparse analysis methods aim to decompose activations into functionally specific units. While Sparse Autoencoders (SAEs) [6, 8] typically target semantic concepts (e.g., "Eiffel Tower"), recent work has focused on identifying functionally specialized neurons. Prior studies have explored techniques ranging from gradient attribution [7, 4] and entropy-based metrics [36] to GPT-based annotation [29]. To overcome the computational cost of supervision, lightweight activation-based approaches—such as average-precision ranking [17] and Neuron-wise Mean Difference (NMD) [13]—have been proposed. These methods have successfully isolated neurons selectively responsible for specific capabilities, including language identification [17, 36], the memory–generalization trade-off [13], and emotional tone [34].

2.3 Confidence: Measurement and Decoupling

Assessment Paradigms. Current research on confidence estimation primarily operates across three dimensions: internal signals, external behavioral indicators, and verbalized self-evaluations. For internal confidence, the most established approach relies on the model's latent states, with Maximum Softmax Probability (MSP) serving as the quintessential baseline [12, 27]. Beyond internal logits, external confidence can be elicited through specific behavioral options, such as providing an "I am not sure" option to allow explicit uncertainty signaling [27]. Finally, directly prompting LLMs for verbalized confidence has become a prevailing paradigm [26, 20]. However, standard verbal elicitations often conflate confidence with accuracy, necessitating a more human-like approach to evaluation.

The Confidence-Performance Decoupling. A core premise of our work is that confidence and performance can be treated as separable variables, a view grounded in recent neuroscience. Theoretically, confidence is not a direct readout of evidence

but an inferential estimate conditioned on an internal self-model [9]. Structurally, this separation is supported by findings that confidence and choice identity are encoded in approximately orthogonal subspaces [39], and that neural populations utilize context signals to shift dynamics without altering synaptic weights [24]. This geometry allows the brain to modulate decisiveness without corrupting the semantic content of the decision. Translating this to LLMs, Xu et al. [41] proposes a decoupling perspective, arguing that verbal estimation should function independently of predictive correctness. In this work, we provide the mechanistic evidence for this decoupling, identifying the specific neural substrate that allows models to regulate action willingness independent of knowledge representation.

3 Empirical Motivation: The Confidence-Performance Decoupling

We evaluate three instruction-tuned models (LLaMA3-8B-IT, Mistral-7B-IT, Qwen3-8B-IT) on MMLU [11]. We construct a role hierarchy (*Expert* → *Student* → *Person* → *Non-Expert*) using a unified template (e.g., *"Now you are a..."*; see Appendix A.1.1). We measure internal confidence via the Maximum Softmax Probability (MSP) of the final answer token. Accuracy is also computed via greedy decoding on final-token logits, yielding a deterministic evaluation. Logit predictions align closely with greedy decoding (see Appendix A.3).

Observation 1: Confidence Decouples from Accuracy. We observe a striking dissociation: role prompts strongly modulate confidence while accuracy remains relatively static. As shown in Figure 2, while accuracy fluctuates non-monotonically, confidence follows a strict, consistent gradient following the role hierarchy from Expert to Non-Expert across all domains. Specific task-level examples confirming this decoupling (e.g., in *College Medicine*) are detailed in Appendix A.2.1. This suggests that role prompts predominantly modulate the model's probabilistic *state of confidence* rather than its underlying knowledge retrieval.

Observation 2: Roles Gate Latent Knowledge via Abstention. To verify if this confidence shift dictates behavior, we employ MMLU-E (adding an *"I am not sure"* option). Table 1 reveals that role identity acts as a decision threshold regulator. The *Non-Expert* induces a conservative state with high abstention (e.g., LLaMA3: 44.8%), while the *Expert* triggers decisiveness (6.9%). Crucially, despite the massive surge in answered questions, *Conditional Accuracy* (Acc_{cond}) remains remarkable stable (69.3% → 67.2%). This stability demonstrates that the "new" answers are not random hallucinations. Instead, they predominantly reflect latent knowledge that was previously suppressed by a conservative confidence gate.

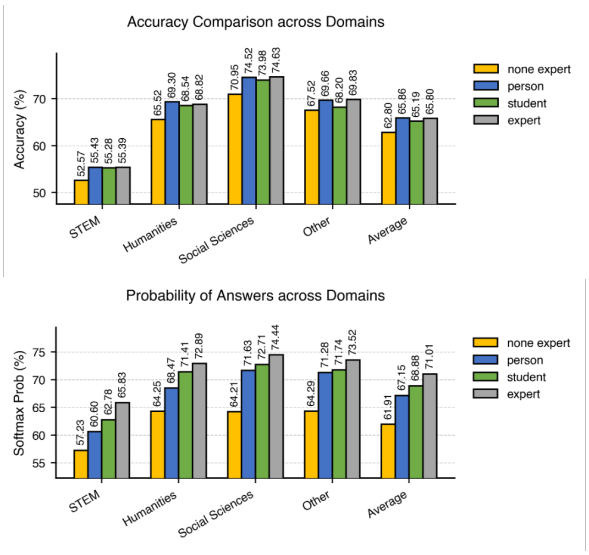


Figure 2: The Decoupling Effect (LLaMA3-8B-IT). While accuracy (*Top*) remains stable across roles, confidence (*Bottom*) follows a strict hierarchy. See Appendix A.2.2 for Mistral and Qwen.

Table 1: Knowledge Unlocking on MMLU-E. Expert prompting sharply reduces abstention (E-ratio) while preserving conditional accuracy (Acc_{cond}), indicating the release of latent knowledge rather than random guessing. All values are percentages (%).

Model	Role	Acc	E-ratio	Acc_{cond}
LLaMA3 8B-IT	Non-Exp.	38.7	44.8	69.3
	Expert	**63.0**	**6.9**	67.2
Mistral 7B-IT	Non-Exp.	21.2	72.7	76.5
	Expert	**50.1**	**24.7**	64.9
Qwen3 8B-IT	Non-Exp.	52.5	29.9	74.9
	Expert	**63.4**	**14.3**	73.9

4 Method: Discovering and Editing Role-Sensitive Neurons (RSNs)

4.1 Overview

Our goal is to isolate the neural substrate of confidence—specifically, the functional ensemble that governs the transition from hesitation to decisiveness. While prior work on activation steering demonstrates that dense, full-layer shifts can modulate behavior, dense vectors inevitably mix confidence signals with semantic content. We propose a finer, neuron-level perspective: identifying a sparse, functionally disentangled control subspace. We focus on the residual stream of the final input token (the decision bottleneck [28]), where integrated context maps to logits. By identifying the neurons that explicitly encode the direction of confidence,

we aim to modulate the model's decisiveness without corrupting its knowledge representation.

4.2 Algorithm: Isolating the Confidence Subspace

We formalize the identification process as a three-step pipeline designed to maximize the Signal-to-Noise Ratio (SNR) of the confidence signal (Figure 1).

Step 1: Collection of Divergent Pairs. To maximize the Signal-to-Noise Ratio (SNR), we filter the dataset for divergent pairs—instances where the prompt change (expert" vs. non-expert") leads to different answer choices. By excluding invariant samples, we ensure the activation difference isolates the specific causal features that tip the model's internal balance toward decisiveness, rather than encoding the shared semantic context of the query.

Step 2: Computation of Role-Induced Shifts. For each divergent pair x_i in our dataset (where $i = 1 \dots N$), we extract the hidden states $\mathbf{h}^{(l)}$ at the final token position. We compute the mean difference [13], which captures the average direction of the activation change:

$$\Delta\mu^{(l)} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{h}^{(l)}_{\text{exp}}(x_i) - \mathbf{h}^{(l)}_{\text{non}}(x_i)\right)$$

This vector $\Delta\mu^{(l)}$ represents the average direction of the role-induced activation shift driven by role conditioning within the layer's activation space.

Step 3: Sparse Filtering (Neuron Isolation). To maximize specificity and minimize interference with unrelated representations, we distill the dense shift vector into a sparse signature. We rank neurons by their absolute shift magnitude $|\Delta\mu_i^{(l)}|$ and retain only the top $\rho\%$ (default $\rho = 0.5\%$), denoted as $\mathcal{K}_l$. The resulting RSN vector isolates the most functionally relevant units while filtering out ambient noise in the dense residual stream:

$$RSNs^{(l)}[i] = \begin{cases} \Delta\mu^{(l)}[i], & \text{if } i \in \mathcal{K}_l, \\ 0, & \text{otherwise.} \end{cases}$$

During inference, we apply the scaled edit $\alpha \cdot RSNs^{(l)}$ to the hidden state of the final input token at each layer. This sparse intervention introduces minimal overhead while preserving the model's capacity.

$$\tilde{h}^{(l)} = h^{(l)} + \alpha \cdot RSNs^{(l)}.$$

4.3 Characterizing the RSN Substrate

We empirically analyze the properties of these extracted neurons, revealing that RSNs constitute a distinct, non-random functional ensemble.

Sparsity (ρ): The Decoupling Evidence. Sweeping the sparsity parameter ρ (Figure 3a) reveals that editing just the top 0.5% of neurons is sufficient to recover expert-level decisiveness while maintaining accuracy. Increasing ρ beyond this threshold yields diminishing returns and eventually degrades performance, suggesting that excessive editing interferes with core semantic features. This efficacy of sparse editing provides circumstantial evidence for the decoupling hypothesis: confidence is encoded in a low-dimensional subspace effectively disentangled from the vast majority of knowledge-storing neurons.

Localization (Layers): The Metacognitive Window. Layer-wise analysis (Figure 3b) shows that role-induced divergence is not uniformly distributed. While early layers (1–10) show high correlation (encoding shared syntax), divergence peaks in the middle layers (11–19). This identifies a middle-layer "metacognitive window"—the locus where the *self-model* asserts influence to inhibit or promote information flow. Consequently, we restrict RSN editing to this region (RSN_{mid}), maximizing stability while minimizing interference (consistent with single-layer baselines, Appendix A.4).

Scaling (α): The Gain Control Mechanism. The scaling factor α acts as a neuromodulatory gain control (Figure 3c). Performance peaks around an optimal range ($\alpha \approx 4$), maximizing the retrieval of correct answers from abstention. Low gain ($\alpha < 2$) is insufficient to overcome the inhibition threshold, while excessive gain ($\alpha > 6$) introduces noise and over-steering. This aligns with the interpretation of RSNs as a gain-modulating signal, providing the necessary "energy" to regulate decisiveness.

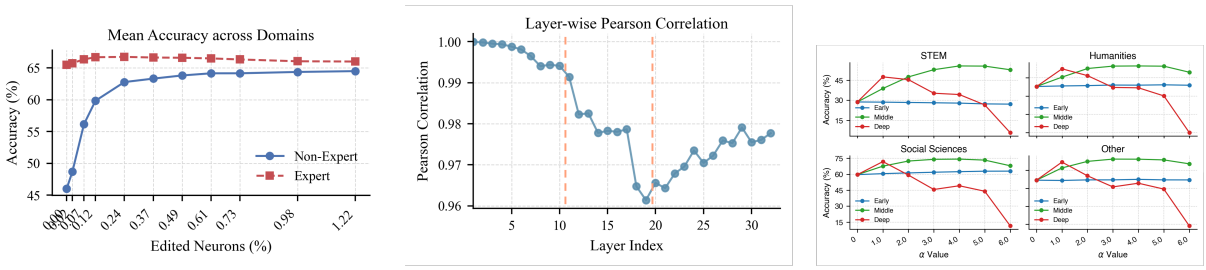
5 Mechanism: Steering the Confidence Switch

Having identified the sparse, role-sensitive subspace, we now move from correlation to causality. In this section, we dissect the neural mechanism of confidence modulation using *LLaMA3-8B-IT* on the *MMLU-E* benchmark as a representative case study. We establish that Role-Sensitive Neurons (RSNs) function not merely as a correlate of confidence, but as a bidirectional *gain control knob* that governs the transition between hesitation and decisiveness. Unless otherwise specified, all causal interventions in this section are applied to the identified *middle-layer RSNs* (RSN_{mid}), as they represent the most stable functional ensemble for confidence steering (see Section 4).

5.1 Causal Control: The Bidirectional Gain Knob

To verify the causal role of RSNs, we perform bidirectional steering (injection and suppression) and compare RSNs against dense and statistical baselines. The results are summarized in Table 2.

Sufficiency: Replicating Decisiveness ($+\alpha$). First, we test whether activating RSNs is sufficient to induce expert-level behavior. As shown in Ta-



(a) Sparsity (ρ) (b) Localization (Layers) (c) Scaling (α)

Figure 3: Characterizing the RSN Substrate (LLaMA3-8B-IT). (a) Sparsity: Performance saturates at low sparsity ($\rho \approx 0.5\%$). (b) Localization: Role-induced divergence peaks in the middle layers (11–19). (c) Scaling: Steering effectiveness peaks at $\alpha \approx 4$, consistent with gain control.

ble 2 (Block II), injecting the sparse middle-layer RSN vector ($\alpha = 4$) into the Non-Expert reduces the abstention rate from 44.77% to 3.73%. This intervention recovers the model's accuracy (63.22%), effectively matching the Expert reference (63.01%) and validating RSNs as a sufficient causal substrate for confidence modulation.

Bidirectional Control: The Continuous Knob (-α). If RSNs represent a continuous mechanistic gain, reversing the signal should induce hesitation. Indeed, applying negative scaling ($\alpha = -3$) produces a strictly opposite effect: the abstention rate spikes to 65.13%—far exceeding the original Non-Expert level. This suppression effect generalizes to the Expert model, increasing its hesitation from 6.87% to 14.64%. This confirms that RSNs function as a prompt-agnostic bidirectional gain mechanism: the neural substrate itself governs the model's decisiveness, regardless of whether the initial state was induced by an expert or non-expert persona.

Table 2: Mechanistic Comparison of Steering Methods (LLaMA3-8B-IT, MMLU-E). (I) Signal Nature: RSNs vs. dense (FV, PCA) and statistical (T-test) baselines. (II) Causal Control: Bidirectional RSN steering demonstrates tunable gain. All values are percentages (%).

		Non-Expert		Expert	
Method	**Config**	**Acc**	**E**	**Acc**	**E**
Original	-	38.65	44.77	63.01	6.87
Random	Sparse	39.70	43.36	63.28	6.54
I. Baselines (Signal Nature)					
PCA (ICV)	Dense	51.47	3.94	60.21	0.14
FV	Dense	62.11	0.10	63.96	0.00
T-test	Sparse	61.81	5.06	65.30	0.17
II. RSN Causal Steering					
$4\,RSN_{mid}$	Sparse	**63.22**	3.73	**65.74**	**0.26**
$-3\,RSN_{mid}$	Sparse	24.95	65.13	58.82	14.64

Necessity: Unmasking Active Suppression. To determine whether RSNs are the necessary physical carrier of role-induced behavior, we perform zeroing ablation (Knockout). We analyze the *Confidence-Capability Gap*—defined as the difference between latent capability (*standard MMLU* accuracy) and expressed behavior (MMLU-E accuracy). Table 3 reveals a striking functional asymmetry. The Non-Expert exhibits a massive confidence-capability gap of 24.15%, indicating active suppression. Crucially, ablating mid-layer RSNs improves behavior, narrowing the gap to 11.03%. This paradoxical improvement implies that the Non-Expert state acts as a "suppression lock" mediated by RSNs. In contrast, the Expert model remains robust to this ablation (the gap remains negligible at ∼2%), proving that RSN removal specifically disengages the suppression mechanism without damaging the core reasoning engine, unlike full-layer knockout which degrades capability globally.

Table 3: Evidence of Active Suppression via RSN Ablation (LLaMA3-8B-IT). Ablating mid-layer RSNs selectively reduces Non-Expert hesitation while preserving accuracy, whereas full-layer ablation degrades latent capability. All values are percentages (%).

Intervention	**MMLU-E**		**MMLU**	**Gap**
	Acc (↑)	**E (↓)**	**Acc (↑)**	
Non-Exp. (Orig.)	38.65	44.77	**62.80**	24.15
Knock(RSN_{mid})	**49.36**	23.25	60.39	11.03
Knock(RSN_{full})	39.10	**14.23**	47.32	8.22
Expert (Orig.)	63.01	6.87	**65.80**	2.79
Knock(RSN_{mid})	**63.04**	**4.02**	65.27	2.23
Knock(RSN_{full})	43.44	5.97	47.88	4.44

5.2 Micro-Dynamics: The Unlocking Mechanism

To understand the mechanics of this performance jump, we analyze prediction transitions and conditional accuracy. This reveals that RSNs function as a *gain control* mechanism—lowering the confidence

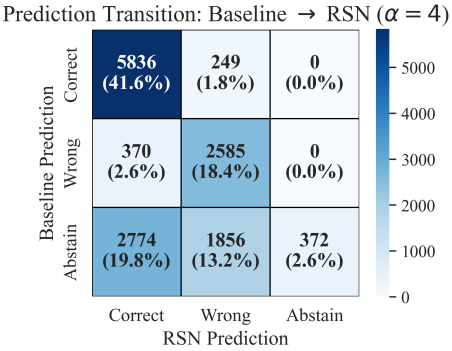


Figure 4: Prediction Transition Matrix (Non-Expert → +RSN). RSN steering induces a dominant *Abstain* → *Correct* transition (19.8%) and a smaller *Abstain* → *Wrong* flow (13.2%), consistent with global threshold lowering.

threshold to release latent knowledge.

Figure 4 visualizes the prediction shifts from the *Non-Expert* baseline to the RSN-steered state ($\alpha = 4$). The dynamics reveal a dominant flow from *Abstain* → *Correct* (19.8% of total samples), confirming that the majority of gains stem from unlocking latent knowledge suppressed by a conservative decision regime. Crucially, this is accompanied by a secondary flow from *Abstain* → *Wrong* (13.2%), whereas the *Correct* → *Wrong* path remains minimal (1.8%).

This pattern confirms that RSNs operate by globally lowering the decision threshold rather than selectively filtering for truth. We further validate this by tracking Conditional Accuracy (Acc_{cond}, see Appendix A.10). The Non-Expert baseline exhibits excessive conservatism ($Acc_{cond} = 69.3\%$). RSN steering shifts this operational point to 65.5%, effectively mirroring the Expert's precision-recall trade-off (67.2%) by trading a marginal drop in precision for a massive expansion in coverage.

5.3 Signal Nature: Efficiency and Specificity

Finally, we analyze the structural properties of the control signal by comparing RSNs against dense baselines—Functional Vectors [38] and PCA-based methods [23]—as well as statistical baselines (T-test) [2]. The results in Table 2 (Block I vs. II) underscore three defining characteristics of the confidence mechanism.

Geometry: Mean Shift vs. Variance (vs. PCA). We find that confidence does not align with the direction of maximum activation variance. The performance gap between PCA-based steering and RSNs supports this distinction. While PCA captures the dominant axes of variation, it fails to isolate the confidence signal. This confirms that confidence is encoded as a *directed mean shift* induced by role conditioning, distinct from the dominant axes of variation.

Efficiency: Minimal Intervention (vs. FV). Standard Functional Vectors (FV) operate on the entire dense layer (100% of neurons). While FV achieves high accuracy, it acts aggressively, driving the E-ratio to near zero. In contrast, RSNs achieve comparable expert-level accuracy by editing only *0.5%* of the neurons. This demonstrates that confidence is carried by a compact and functionally specialized subspace, rather than being diffusely distributed across the layer.

Specificity: Magnitude vs. Significance (vs. T-test). Comparing RSNs against the T-test baseline offers insight into the signal's composition. T-test selects neurons based on strict statistical consistency (low variance across samples), whereas RSNs select based on the absolute magnitude of divergence. The slight performance advantage of RSNs suggests that strict statistical stability at the level of individual neurons is not a necessary condition for effective control. Instead, what matters is whether the aggregate mean shift is sufficient to reliably bias the representation along the confidence dimension.

Together, these results characterize confidence as a sparse, directional, and population-level control signal rather than a generic statistical artifact.

5.4 Validation: Internal State Dynamics

Does RSN injection genuinely modulate the model's internal belief state, or merely disable refusal heuristics? To distinguish these possibilities, we validate the mechanism using two complementary metrics: internal confidence (MSP) and verbalized self-evaluation.

Internal Confidence (MSP). We track the model's latent belief state (p^*) under causal intervention. To verify generalization beyond the "Expert" template, we introduce semantic anchors—explicitly *Confident* and *Unconfident* prompts (see Appendix A.1.2). As shown in Figure 5, RSN injection induces a consistent monotonic increase in p^* across all conditions. This cross-prompt consistency is consistent with RSNs regulating a fundamental internal signal, effectively "boosting" latent confidence regardless of the initial framing.

Alignment with Verbalized Confidence. We further verify that this internal shift aligns with the model's explicit self-assessment. On MMLU-Pro, RSN steering consistently raises and stabilizes self-evaluated confidence scores (0–9 scale). This synchronization between implicit logits and explicit generation suggests that the gain control effect is also reflected in the model's verbalized outputs (see Appendix A.12 for detailed distribution analysis).

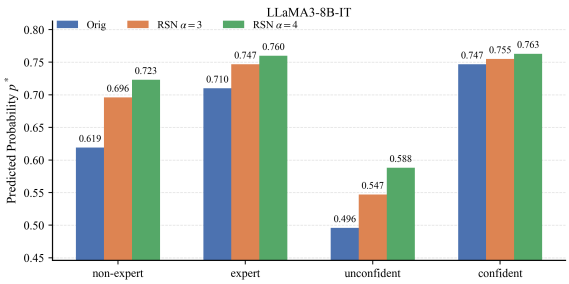


Figure 5: Confidence Amplification under RSN Editing. RSN intervention consistently increases the predicted probability (p^*) across role-based and explicit confidence prompts in *LLaMA3-8B-IT*, indicating a generalized boost in latent certainty. See Appendix A.11 for consistent results on *Mistral-7B-IT* and *Qwen3-8B-IT*.

6 Generalization and Boundary Analysis

Having established the mechanistic nature of RSNs, we evaluate their robustness across unseen roles, neutral settings, and boundary regimes where increased confidence does not imply factual correctness.

6.1 Cross-Domain Generalization: New Roles and Formats

We first test whether mid-layer RSNs generalize to scientific (ARC Challenge [5]) and commonsense reasoning (CSQA [35]) without retraining. We use the MMLU-E template format but with generic roles without task-specific descriptions (e.g., *Person*, *Student*, *Level-n Expert*, see Appendix A.1.3). As shown in Table 4, RSN editing on LLaMA3-8B-IT demonstrates strong transferability. Weak roles (e.g., *Non-Expert*) exhibit the most dramatic recovery, with accuracy surging over 30 points as abstention vanishes. Crucially, the graded reduction in E-ratios across generic levels confirms that RSNs function as a task-agnostic, continuous gain controller, rather than relying on task-specific semantic cues. Consistent generalization on Mistral and CSQA-E is reported in Appendix A.7.

6.2 Standalone Confidence Control in Neutral Settings

Can RSNs replace prompt engineering? We evaluate direct intervention in a *Neutral* condition (no role prompts). Table 5 reveals a clear *Reasoning-Factuality Trade-off*. Positive injection (α=4) consistently amplifies reasoning performance (e.g., GPQA +12.1%), suggesting RSNs promote weak but correct latent chains. However, this gain control is distinct from epistemic correctness: excessive boosting lowers factual accuracy, while negative steering ($\alpha = -4$) acts as a "caution filter," improving truthfulness. This confirms that RSNs modulate the willingness to assert rather than directly improving the verification of truth. (See Appendix A.13 for Mistral's distinct dynamics).

Table 4: Cross-Domain Generalization on ARC-E. RSN editing (L_{11-19}, α=4) on LLaMA3-8B-IT consistently reduces abstention across generic roles. All values are percentages (%).

Role	Accuracy		E-ratio	
	Orig.	Edit.	Orig.	Edit.
Non-Exp.	47.22	**80.46**	43.90	**0.46**
Person	52.63	**81.31**	36.49	**0.12**
Student	60.39	**81.27**	28.80	**0.15**
Expert	80.58	**82.16**	3.17	**0.04**
L0 Exp.	60.39	**82.01**	29.50	**0.19**
L1 Exp.	80.31	**82.30**	3.98	**0.08**
L2 Exp.	81.31	**82.28**	2.20	**0.08**
L3 Exp.	81.66	**82.24**	1.66	**0.08**

6.3 Origin of RSNs: Cross-Model Transferability

Finally, we investigate whether RSNs are novel circuits created by instruction tuning (IT), or latent mechanisms already present in pre-training. We perform a cross-model transfer experiment on *MMLU-E*, directly applying RSNs extracted from the IT model to the corresponding *Base* model without retraining. As shown in Table 6, injecting the IT-RSN vector ($\alpha = 4$) into *LLaMA3-Base* (Non-Expert) reduces abstention from 61.17% to 7.39% and more than doubles accuracy (20.67% $\rightarrow$ 47.47%), substantially outperforming the Base model's own extracted vector (29.11%).

This result demonstrates that the confidence control mechanism is *indigenous to pre-training* but initially unrefined. The Base model's strong response to the transferred IT vector supports the existence of an underlying latent structure, while the superior performance of IT-RSNs indicates that instruction tuning acts as a *signal sharpener*, isolating and amplifying a precise gain-control direction from noise.

7 Theoretical Implications

Our findings establish RSNs as a specialized neural subspace for regulating mechanistic confidence. Beyond immediate performance gains, these neurons offer critical insights into the nature of LLM cognition.

7.1 The "Digital Dopamine" Hypothesis

Neuroscience posits that biological gain control mechanisms (e.g., dopaminergic pathways in the NAcc) regulate decisiveness not merely by altering sensory evidence, but by modulating the commitment to a choice [16]. Specifically, high dopamine levels provide the "energetic" motivation to over-

Table 5: Standalone RSN Control in Neutral Settings. RSN editing improves reasoning benchmarks but reveals a trade-off with factuality.

Model	Cond.	Reasoning Benchmarks (Acc %)					Factuality (Acc %)		
		MMLU	MMLUpro	GPQA	AR-LSAT	LogiQA	TQA-MC1	TQA-MC2	FACTOR
LLaMA3	Orig	**67.4**	36.1	31.9	23.2	54.5	51.0	59.9	71.6
	$\alpha = +4$	66.4	**37.8**	**32.8**	**23.5**	**55.3**	46.0	56.6	68.3
	$\alpha = -4$	66.9	33.8	31.6	22.5	54.3	**51.3**	**61.4**	**72.8**
Qwen3	Orig	71.7	41.1	33.4	25.6	66.8	**68.1**	76.6	75.8
	$\alpha = +4$	**72.4**	**43.7**	**35.6**	**26.1**	**67.5**	66.7	**77.0**	**77.0**
	$\alpha = -4$	67.7	35.6	30.3	25.2	62.5	65.2	73.2	69.4

Table 6: Cross-Model Transfer. IT-extracted RSNs effectively steer the Base model, outperforming its own vectors in the *non-expert* setting.

Setting (Non-Expert)	LLaMA3-8B-Base		Mistral-7B-Base		Qwen3-8B-Base	
	Acc (%)	E (%)	Acc (%)	E (%)	Acc (%)	E (%)
Original Baseline	20.67	61.17	3.30	94.86	57.02	18.25
+ Base RSN (Self)	29.11	42.59	8.49	84.35	67.24	0.81
+ **IT RSN (Transfer)**	**47.47**	**7.39**	**34.33**	**23.58**	**68.01**	**0.06**

come internal inhibition and prevent "change-of-mind quitting" during difficult decisions.

Our results suggest a striking parallel in LLMs: RSNs function analogously to "Digital Dopamine." We use this term metaphorically to denote gain modulation. RSNs provide the computational equivalent of "motivation to engage": simply increasing the scaling factor α suppresses the model's tendency to abstain, effectively pushing the internal state past the activation threshold required for decisive action. This supports the view that LLMs separate the *representation of truth* from the willingness to act on it, employing a functionally dissociable coding scheme for confidence.

7.2 Alignment as Signal Sharpening

The "Superficial Alignment Hypothesis" [43] posits that instruction tuning (SFT) primarily exposes latent capabilities rather than injecting new skills. Recent work further demonstrates that base models possess strong latent reasoning potential that is masked by flat output distributions, and can be unlocked by sharpening the sampling process [3].

Our cross-model transfer experiments (Section 6.3) provide the mechanistic substrate for these theories. We show that the "sharpening" mechanism is structurally localized in RSNs. The base model already possesses these latent functional structures, but they operate in a low-gain, noisy regime. SFT functions as a *Gain Tuning* process, systematically biasing the RSN knob to a high-gain setting. This explains why transferring the "sharpened" IT-RSN vector back to the Base model instantly unlocks decisiveness without additional training. While this gain tuning perspective explains how alignment sharpens decisiveness, it also underscores that confidence modulation alone does not guarantee epistemic correctness, a limitation we discuss in detail in Section 8.

8 Conclusion

In this work, we identify and characterize *Role-Sensitive Neurons (RSNs)*, a specialized subnetwork governing the mechanistic transition from hesitation to action. By locating these neurons through the lens of role-playing, we demonstrate that LLM confidence is not merely a statistical readout of evidence, but a controllable internal variable regulated by a specific *gain control neural subspace*.

Our experiments validate a clear *Confidence–Performance Decoupling*: RSNs operate largely independently of the model's factual knowledge. Intervening on this substrate can "unlock" latent capabilities suppressed by weak self-models (e.g., LLaMA3), or induce unwarranted certainty in knowledge-deficient regimes (e.g., Mistral). Cross-model transfer further supports the view that this mechanism is indigenous to pre-training, with instruction tuning acting as a "signal sharpener" rather than creating the capability *de novo*.

These findings bridge behavioral alignment with mechanistic interpretability, characterizing "becoming an expert" as a gain modulation process. However, the dissociation between gain and knowledge highlights a critical safety imperative. *Future work* should investigate temporal confidence dynamics during multi-step reasoning and extend this analysis to *Mixture-of-Experts (MoE)* architectures—using expert routing as a natural testbed to compare implicit neural gain with explicit architectural gating. Ultimately, RSNs offer neuron-level monitoring signals to distinguish calibrated certainty from artificial overconfidence.

Limitations

While our work offers a mechanistic explanation for confidence regulation in LLMs, we acknowledge several limitations that frame the scope for future research.

Model Scale and Architecture. Our analysis primarily focuses on dense models in the 7B–8B parameter range (LLaMA3, Mistral, Qwen). While these models exhibit consistent RSN behaviors, it remains an open question whether similar gain control circuits exist in Mixture-of-Experts (MoE) architectures or significantly larger models (e.g., 70B+), where confidence mechanisms might be more distributed or redundant.

Task Scope. We evaluated RSNs predominantly on reasoning and knowledge-intensive benchmarks (MMLU, ARC, CSQA). The role of gain control in open-ended creative generation or long-context retrieval tasks remains unexplored. It is possible that "confidence" in creative writing relies on distinct neural substrates from the epistemic confidence required for factual QA.

Diagnosis vs. Mitigation. Finally, our work identifies the mechanism of "unwarranted certainty" but does not propose an algorithmic defense. While we suggest RSN probing as a monitoring tool, developing training objectives (e.g., specific regularizers) to automatically penalize RSN over-activation during hallucinations is a critical next step for AI safety.

Ethics Statement

Our research is primarily mechanistic in nature, focusing on the internal activation patterns of large language models. We utilize only publicly available datasets (e.g., MMLU, ARC) and open-weight models (e.g., LLaMA, Mistral, Qwen) in strict accordance with their original licenses and intended use. No human subjects were involved in our experiments, and no new sensitive personal data was collected. While we discuss the potential for "unwarranted certainty," our findings are intended as a diagnostic contribution to AI safety and model calibration, rather than a tool for facilitating the generation of deceptive or harmful content.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[2] Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2024. The llm language network: A neuroscientific approach for identifying causally task-relevant units. *arXiv preprint arXiv:2411.02280*.

[3] Anonymous. 2025. Reasoning without training: Your base model is smarter than you think. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

[4] Lihu Chen, Adam Dejl, and Francesca Toni. 2025. Identifying query-relevant neurons in large language models for long-form texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23595–23604.

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

[6] 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*.

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

[8] Boyi Deng, Yu Wan, Yidan Zhang, Baosong Yang, and Fuli Feng. 2025. Unveiling language-specific features in large language models via sparse autoencoders. *arXiv preprint arXiv:2505.05111*.

[9] Stephen M Fleming. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268.

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

[11] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

[12] Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

[13] Ko-Wei Huang, Yi-Fu Fu, Ching-Yu Tsai, Yu-Chieh Tu, Tzu-Ling Cheng, Cheng-Yu Lin, Yi-Ting Yang, Heng-Yi Liu, Keng-Te Liao, Da-Cheng Juan, and 1 others. 2025. Neuron-level differentiation of memorization and generalization in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16077–16091.

[14] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

[15] Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a double-edged sword: Enhancing the zero-shot reasoning by ensembling the role-playing and neutral prompts. *arXiv e-prints*, pages arXiv–2408.

[16] Adrina Kocharian, A David Redish, and Patrick E Rothwell. 2025. Individual differences in decision-making shape how mesolimbic dopamine regulates choice confidence and change-of-mind. *Nature Neuroscience*, 28(9):1883–1896.

[17] Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971.

[18] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113.

[19] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Jiaming Zhou, and Haoqin Sun. 2024. Self-prompt tuning: Enable autonomous role-playing in llms. *arXiv preprint arXiv:2407.08995*.

[20] Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. 2024. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334.

[21] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

[22] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3622–3628.

[23] Sheng Liu, Haotian Ye, Lei Xing, and James Y Zou. 2024. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *International Conference on Machine Learning*, pages 32287–32307. PMLR.

[24] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.

[25] Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66.

[26] Yudi Pawitan and Chris Holmes. 2025. Confidence in the reasoning of large language models. *Harvard Data Science Review*, 7(1):2644–2353.

[27] Benjamin Plaut, Khanh Nguyen, and Tu Trinh. 2024. Softmax probabilities (mostly) predict large language model correctness on multiple-choice q&a. *CoRR*.

[28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[29] Daking Rai and Ziyu Yao. 2024. An investigation of neuron activation as a unified lens to explain chain-of-thought eliciting arithmetic reasoning of llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7174–7193.

[30] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

[31] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522.

[32] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-context impersonation reveals large language models' strengths and biases. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[33] Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

[34] Brenden Smith, Dallin Baker, Clayton Chase, Myles Barney, Kaden Parker, Makenna Allred, Peter Hu, Alex Evans, and Nancy Fulda. 2024. The mysterious case of neuron 1512: Injectable realignment architectures reveal internal characteristics of meta's llama 2 model. *arXiv preprint arXiv:2407.03621*.

[35] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

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

[37] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631.

[38] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

[39] Miguel Vivar-Lazo and Christopher R Fetsch. 2025. Neural basis of concurrent deliberation toward a choice and confidence judgment. *Nature Neuroscience*, pages 1–12.

[40] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan

Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

[41] Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. 2025. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms. *arXiv preprint arXiv:2506.00582*.

[42] Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. *arXiv preprint arXiv:2104.06598*.

[43] Chunting Zhou, Pengfei Liu, Puxuan Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Supplementary Figures and Tables

A.1 Prompt Templates

We employ distinct prompt templates for standard evaluation, confidence modulation, uncertainty estimation (MMLU-E), and self-evaluation.

A.1.1 Standard Task Template

This template is used for all experiments on the original datasets (MMLU, MMLU-Pro, PubMedQA, MedQA, GPQA, etc.) where the "I am not sure" option is absent. `[task]` refers to the specific subject name (e.g., *anatomy*). For datasets without domain-specific labels, generic roles are used.

Role Instantiations:

- `Now you are a [task] expert` (MMLU / MMLU-Pro)
- `Now you are a non [task] expert` (MMLU / MMLU-Pro)
- `Now you are an expert / a non expert` (General tasks)
- `[no role prefix]` (Neutral condition)

```
Would you answer the following
question with A, B, C or D?
Question: What is a normative
perspective?
A) The way things are.
B) The way things are going to be.
C) The way things should be.
D) The way things were.
[Role prefix appears here if applicable, e.g.,
Now you are a (non) [task] expert]
Your answer among A, B, C or D is:
```

A.1.2 Confidence Modulation Template

To directly manipulate the model's internal confidence level without changing the persona, we replace the role phrase with explicit confidence descriptors.

- Confident: `Now you are confident.`
- Unconfident: `Now you are unconfident.`

```
Would you answer the following
question with A, B, C or D?
Question: What is a normative
perspective?
A) The way things are.
B) The way things are going to be.
C) The way things should be.
D) The way things were.
Now you are confident.
Your answer among A, B, C or D is:
```

A.1.3 MMLU-E / ARC-E / CSQA-E Template

To capture uncertainty, we extend the standard template by adding *E) I am not sure*. We also add the adjective "honest" to the role description to emphasize sincerity in abstention.

Role Instantiations:

- `Now you are an honest [task] expert` (MMLU-E)
- `Now you are an honest non [task] expert` (MMLU-E)
- `Now you are an honest person` (MMLU-E / ARC-E / CSQA-E)
- `Now you are an honest level [n] expert` (ARC-E / CSQA-E)

```
Would you answer the following
question with A, B, C, D or E?
Question: What is a normative
perspective?
A) The way things are.
B) The way things are going to be.
C) The way things should be.
D) The way things were.
E) I am not sure.
[Role prefix appears here, e.g., Now you are
an honest (non) anatomy expert]
Your answer among A, B, C, D or E is:
```

A.1.4 Reasoning Willingness Self-Evaluation

To measure mechanistic confidence via self-report, we prompt the model to score its readiness to reason on a 0–9 scale. No explanation or calibration is provided.

```
Here is a question: What is a
normative perspective?
A) The way things are.
B) The way things are going to be.
C) The way things should be.
D) The way things were.
[Condition description, e.g., Now you are a
non [task] expert]
When dealing with the question,
your self-evaluation of reasoning
willingness from 0 to 9 is:
```

A.2 Behavioral Analysis

A.2.1 Task-level Breakdown of Role Effects

Table 7 provides specific examples of the dissociation between accuracy and confidence across different models and tasks.

A.2.2 Cross-Model Consistency of Role Effects

Figure 6 presents results for *Mistral-7B-IT* and *Qwen3-8B-IT*, complementing the LLaMA3-8B-IT plots shown in the main paper. Across both models, we observe a highly consistent pattern: confidence follows a clear hierarchical gradient (`expert` > `student` > `person` > `non-expert`), whereas accuracy remains largely unchanged across domains. This confirms that the dissociation between internal certainty and output correctness is not model-specific, but a general effect of role conditioning.

A.3 Validation: Correlation between Generation and Logits

To validate the reliability of our evaluation protocol, we compared results computed from logits

Table 7: Role-conditioned accuracy and confidence (in %) across tasks and models. Note that while confidence strictly increases with stronger roles, accuracy does not necessarily follow the same trend.

Task	N-Exp.	Person	Student	Exp.
Accuracy				
College Medicine (LLaMA3)	**69.9**	67.6	67.5	69.4
Business Ethics (Mistral)	**62.0**	61.0	**62.0**	**62.0**
College Chem. (Qwen)	**56.0**	53.0	55.0	55.0
Confidence (MSP)				
College Medicine (LLaMA3)	68.1	69.0	70.0	**71.3**
Business Ethics (Mistral)	71.5	75.9	77.4	**78.3**
College Chem. (Qwen)	76.7	77.7	79.0	**79.7**

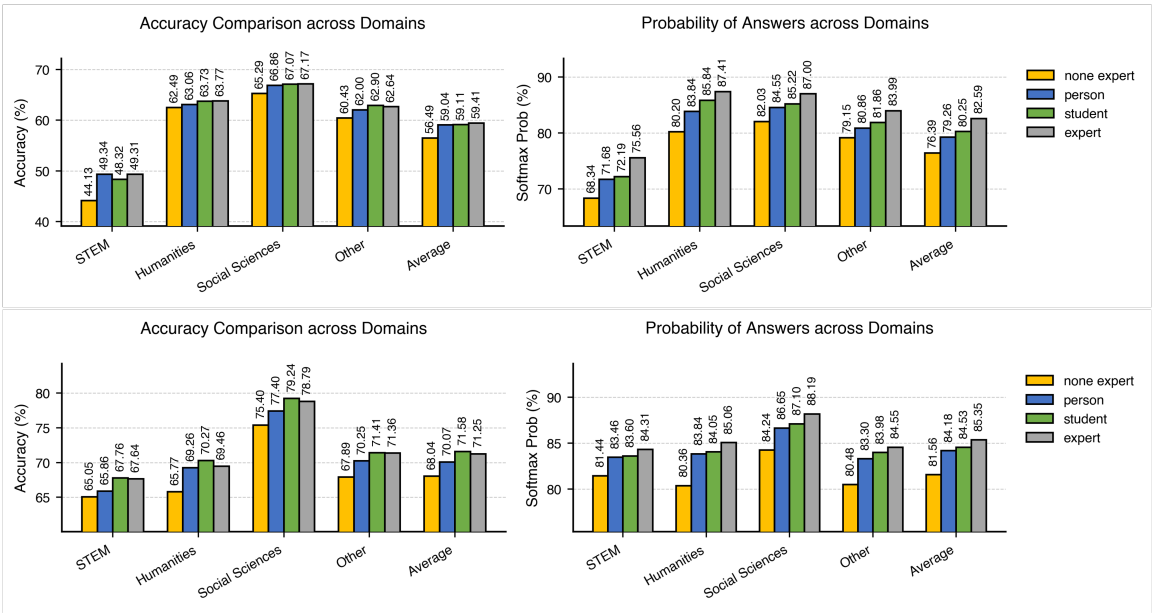


Figure 6: **Cross-Model Consistency.** Accuracy (left) vs. confidence (right) across domains under different role prompts for *Mistral-7B-IT* (top) and *Qwen3-8B-IT* (bottom). Both models replicate the same trend observed in LLaMA3-8B-IT, highlighting the universality of role-induced confidence modulation.

with those obtained from greedy generation. Specifically, we evaluated both accuracy and abstention (E-ratio) across tasks and roles for LLaMA3-8B-IT and Mistral-7B-IT. As shown in Figure 7 (Left), the two evaluation methods yield nearly identical results on accuracy, with Pearson's R values exceeding 0.96 across roles and models. Similarly, Figure 7 (Right) shows that abstention rates computed from logits strongly correlate with those from generation. These results confirm that logits-based evaluation is a faithful and computationally efficient proxy for generation-based metrics.

A.4 Localization: Single-Layer Functional Vector Editing

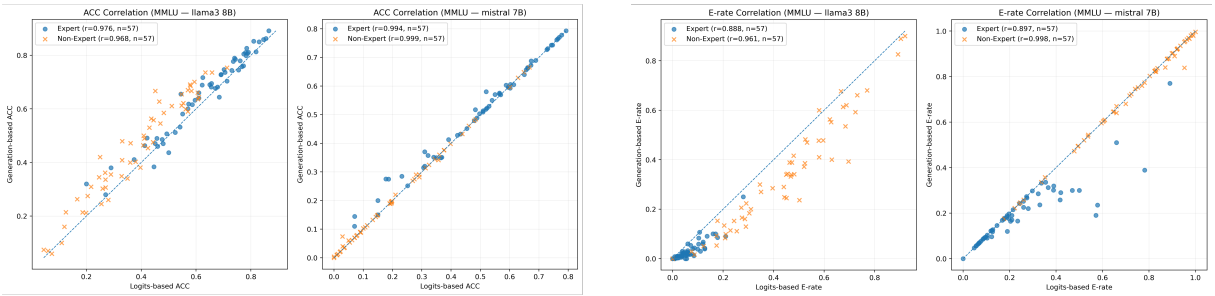
To further localize role-sensitive transformations, we applied Functional Vector (FV) editing (using *all neurons* in a single layer, $\alpha = 1$) to LLaMA3-8B-IT across four representative MMLU tasks ("college computer science," "US foreign policy," "management," and "jurisprudence"), each from a different domain.

As shown in Figure 8, editing effects begin to emerge around layer 11, rise steeply through the middle layers, and gradually saturate in deeper layers. This indicates that the middle layers (11–19) are the main locus where role-sensitive representations are encoded. This finding provides convergent evidence for our refined RSN design, which focuses on middle-layer edits to maximize steering efficiency while reducing interference.

A.5 Domain-wise Effect of Neuron Editing Sparsity

To further examine domain-specific variation, we plot accuracy under the *expert* and *non-expert* roles for each MMLU domain: STEM, Humanities, Social Sciences, and Other.

Across all domains (Figure 9), the *non-expert* role consistently benefits from sparse neuron edits up to $\rho \approx 0.5\%$, after which performance saturates or slightly declines. The *expert* role remains relatively



(a) Accuracy Correlation

(b) Abstention (E-ratio) Correlation

Figure 7: Validation of Logit-Based Evaluation. Correlation between logit-based (x-axis) and generation-based (y-axis) metrics on MMLU for LLaMA3-8B-IT and Mistral-7B-IT. Each point represents a specific task-role pair. The high correlation ($R > 0.96$) justifies our use of logits for efficient evaluation.

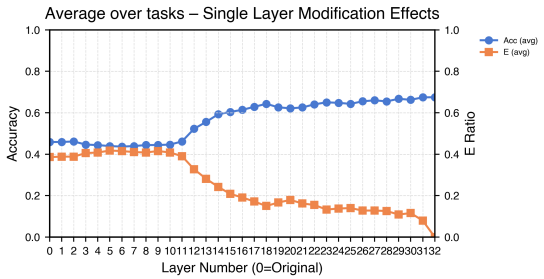


Figure 8: Layer-wise Steering Effects. Average performance over four MMLU tasks on *LLaMA3-8B-IT* using single-layer functional vector editing ($\alpha = 1$). Role-sensitive effects (increased accuracy, decreased abstention) emerge distinctly around layer 11 and peak in the middle layers, corroborating our strategy to target the middle-layer RSN subspace.

stable, confirming that role-conditioned behavior can be selectively enhanced without degrading well-calibrated expert reasoning.

A.6 Layer-wise Correlation For Mistral and Qwen3

To examine whether the mid-layer concentration of role-sensitive signals generalizes across architectures, we compute the layer-wise Pearson correlation between "expert" and "non-expert" hidden states in both *Mistral-7B-IT* and *Qwen3-8B-IT*.

As shown in Figure 10, both models exhibit a similar three-phase pattern to *LLaMA3-8B-IT*: (1) early layers show nearly identical activations, (2) middle layers diverge sharply (Mistral: layers 14–21; Qwen3: 17–25), and (3) deeper layers partially recover. This cross-model consistency suggests that role-specific transformations are robustly localized to middle layers across diverse transformer architectures.

A.7 Extended Cross-Domain Generalization Results

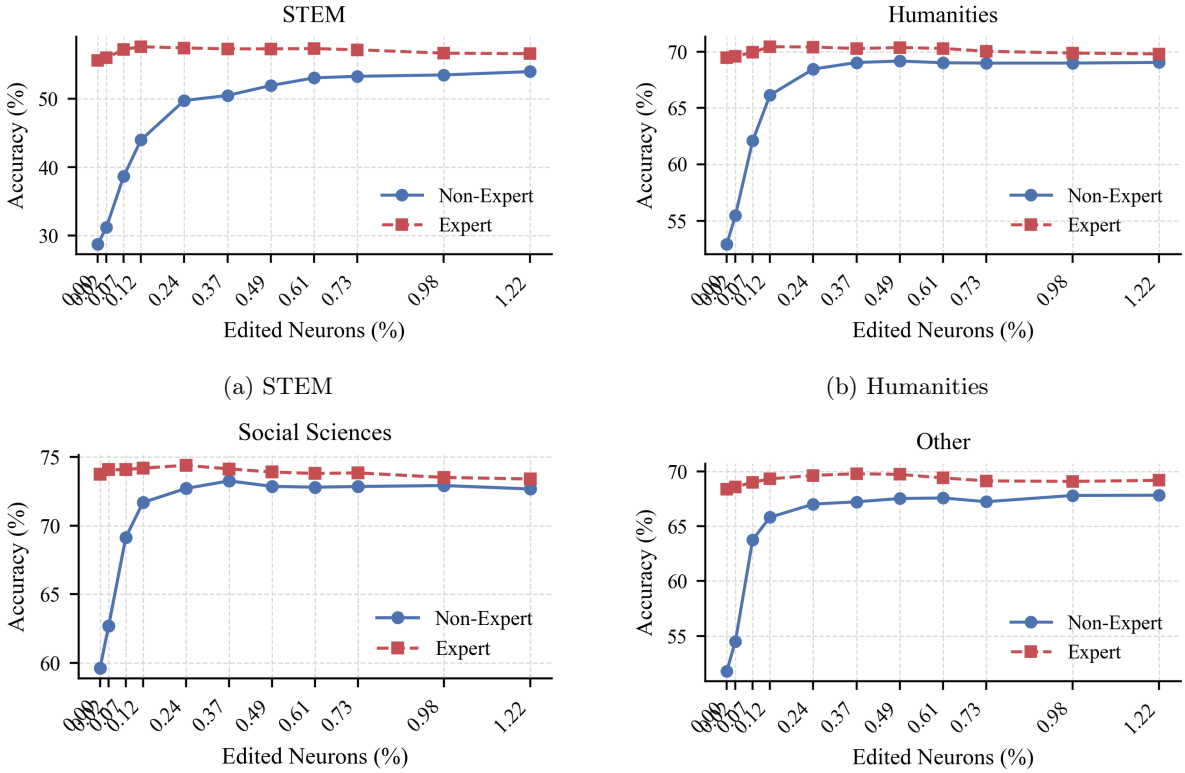
This appendix provides supporting evidence for the cross-domain robustness of RSNs discussed in Section 6.1. We present two complementary analyses: (1) the generalization of Mistral-7B-IT on the *ARC Challenge*, and (2) the performance of both LLaMA3 and Mistral on *CommonSenseQA* (CSQA).

Mistral-7B-IT on ARC-E. Table 8 details the performance of Mistral-7B-IT on ARC-E using generic roles. Consistent with LLaMA3, Mistral exhibits a strong response to RSN editing (L_{14-22}, α=4). In the *Non-Expert* role, accuracy doubles (22.55% $\rightarrow$ 46.76%) as the model overcomes its extreme conservatism, confirming that the gain control mechanism functions across different model architectures.

CommonSenseQA Analysis (CSQA-E). Table 9 reports results on *CommonSenseQA* (CSQA-E) for both models. Consistent with ARC findings, RSN editing improves accuracy and calibration across all roles. The effect is most pronounced for weaker roles (*non-expert*, *student*), where abstention is drastically reduced. For stronger expert roles, RSNs continue to reduce residual uncertainty (E-ratio), verifying that RSNs generalize not only across tasks (Scientific vs. Commonsense) but also across diverse role phrasings.

A.8 Generalization Across Prompt Formats

To verify that RSN-based steering is not tied to specific prompt wordings, we evaluate model behavior under differing prompt formulations. All experiments use the abstention-enabled variant (with "E) I am not sure"), and apply the same refined middle-layer RSNs from the main experiments ($\text{RSNs}_{[11,19]}$ for LLaMA3-8B-IT, α=3 or 4)[cite: 844, 845, 846].



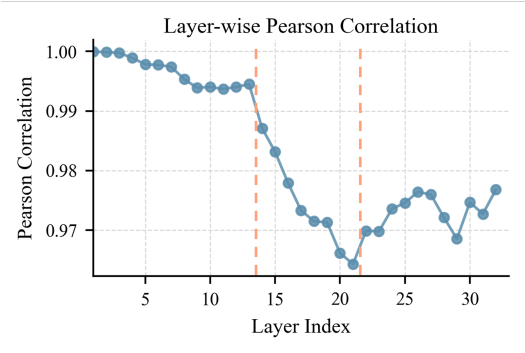
(a) STEM

(b) Humanities

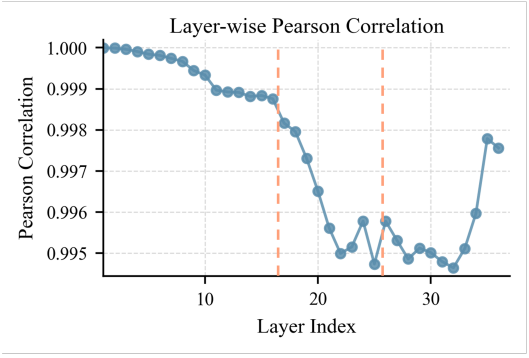
(c) Social Sciences

(d) Other

Figure 9: Domain-wise Editing Sparsity. Accuracy under expert and non-expert roles across varying neuron sparsity levels (ρ). The overall pattern mirrors the mean trend (Figure 3c), with non-expert gains consistently saturating around $\rho \approx 0.5\%$ across all domains, while expert performance remains steady.



(a) Mistral-7B-IT: Divergence peaks at layers 14–21.



(b) Qwen3-8B-IT: Divergence peaks at layers 17–25.

Figure 10: Cross-Model Layer-wise Correlation. Pearson correlation between "expert" and "non-expert" hidden states. Both models replicate the mid-layer divergence pattern observed in *LLaMA3-8B-IT* (Figure 3b).

Table 8: Generalization on Mistral-7B-IT (ARC-E). RSN editing (L_{14-22}, α=4) significantly reduces abstention and recovers accuracy across generic roles, replicating the mechanism observed in LLaMA3.

Role	Accuracy		E-ratio	
	Orig.	Edit.	Orig.	Edit.
Non-Exp.	22.55	**46.76**	73.86	**44.71**
Person	45.48	**52.51**	44.63	**36.72**
Student	48.03	**59.96**	41.93	**29.38**
Expert	63.86	**80.54**	19.69	**3.01**
L0 Exp.	37.57	**60.66**	53.98	**29.42**
L1 Exp.	64.86	**80.19**	16.95	**4.25**
L2 Exp.	65.37	**81.39**	15.87	**2.24**
L3 Exp.	65.52	**81.54**	15.79	**1.85**

Table 9: Performance on CommonSenseQA (CSQA-E). RSN editing ($\alpha = 4$) consistently unlocks performance in weaker roles while maintaining high accuracy in expert roles across both LLaMA3-8B-IT and Mistral-7B-IT.

Model	Role	Accuracy (%)		E-ratio (%)	
		Orig.	Edit.	Orig.	Edit.
LLaMA3	Non-Exp.	62.90	**72.73**	14.99	**0.25**
	Person	47.50	**72.65**	36.36	**0.57**
	Student	53.48	**72.73**	28.01	**0.08**
	Expert	71.91	**73.96**	5.00	**0.16**
	L0 Exp.	52.66	**72.73**	30.71	**0.25**
	L1 Exp.	72.56	**72.81**	3.28	**0.08**
	L2 Exp.	**74.28**	73.46	1.31	**0.08**
	L3 Exp.	**74.20**	72.97	1.56	**0.08**
Mistral	Non-Exp.	30.79	**57.25**	61.43	**19.00**
	Person	57.90	**66.01**	22.11	**4.18**
	Student	57.82	**66.18**	22.60	**4.91**
	Expert	60.28	**67.90**	20.07	**1.80**
	L0 Exp.	47.58	**63.72**	37.59	**6.80**
	L1 Exp.	62.16	**66.75**	16.05	**1.97**
	L2 Exp.	62.00	**67.32**	16.54	**2.05**
	L3 Exp.	61.67	**67.16**	16.63	**2.21**

A.8.1 Vanilla Instruction Format

We first replace the standard role template with a plain instruction-style prompt, stripping away descriptive framing:

```
What is a normative perspective?
A) The way things are.
...
E) I am not sure.
```
As an honest (non) [task] expert, answer:

As shown in Table 10 (Panel A), RSN editing substantially improves both accuracy and decisiveness under this minimal instruction format. Even without contextual hints, the refined RSNs—derived from a different prompt formulation—yield consistent gains across all domains.

A.8.2 Chat-Template Format

Next, we evaluate RSN steering under a chat-template format, where role instructions are injected into the system message:

```
System: Now you are a (non) [task]
expert.
User: Would you answer the following
question with A, B, C, D or E?
Question: ...
Your answer among "A, B, C, D, E" is:
```

As shown in Table 10 (Panel B), RSN steering generalizes robustly to chat-formatted contexts. Middle-layer RSNs produce large accuracy gains with near-zero abstention, indicating that the steering effect persists under dialogue-style prompting.

Taken together, these findings demonstrate that RSNs capture latent role representations intrinsic to model computation rather than superficial textual cues.

Table 10: Robustness to Prompt Formats (LLaMA3-8B-IT). RSN steering effectively modulates confidence across (A) Vanilla Instruction and (B) Chat Template formats. The mechanism generalizes beyond the specific wording used for neuron discovery[cite: 951, 954].

Condition	STEM	Humanities	Social Sci.	Other	Avg. Acc	E (%)
Panel A: Vanilla Instruction Format						
Non-Expert (Baseline)	17.9	25.4	37.2	29.9	26.4	53.8
+ RSN $\alpha = 3$	31.0	40.0	45.9	41.3	38.5	23.1
+ RSN $\alpha = 4$	36.3	43.3	46.2	44.4	41.8	8.3
Expert (Baseline)	33.8	48.8	51.8	44.5	43.4	22.9
+ RSN $\alpha = 3$	40.1	53.4	51.8	49.9	47.8	6.0
+ RSN $\alpha = 4$	40.4	49.9	50.1	48.0	46.3	2.0
Panel B: Chat-Template Format						
Non-Expert (Baseline)	40.2	62.3	63.5	61.2	54.9	18.0
+ RSN $\alpha = 3$	51.7	66.6	67.8	64.3	61.4	1.0
+ RSN $\alpha = 4$	50.7	65.6	68.1	63.3	60.7	0.3
Expert (Baseline)	49.0	64.1	67.1	63.9	59.6	5.9
+ RSN $\alpha = 3$	52.7	66.5	68.2	64.4	61.8	0.3
+ RSN $\alpha = 4$	50.8	65.7	68.3	63.5	60.8	0.1

A.9 Benchmark Descriptions

We evaluate the effects of RSN editing on other representative benchmarks besides MMLU, that do not include an explicit abstention option. These datasets cover a broad spectrum of cognitive demands, including *(i) general reasoning ability* (MMLU, MMLU-Pro, GPQA), *(ii) pure logical inference* (AR-LSAT, LogiQA), *(iii) factual recall and truthfulness* (TruthfulQA, FACTOR).

- **MMLU-Pro** [40]: a more challenging extension of MMLU that emphasizes multi-step and compositional reasoning, featuring up to 9–10 options per question and improved robustness to prompt variation. We use the official `test` split for evaluation.
- **GPQA** [30]: a graduate-level science benchmark spanning physics, biology, and chemistry, designed to test conceptual understanding and cross-disciplinary reasoning. We use the `main` and `diamond` subsets, merging them into a unified pool and constructing 4-option multiple-choice questions by combining the correct answer with three provided distractors in randomized order.
- **AR-LSAT** [42]: a law-domain reasoning dataset modeled after LSAT analytical reasoning, evaluating abstract reasoning and analogical logic under structured constraints. We use the full dataset released on GitHub.
- **LogiQA** [22]: a reading-comprehension–style logical reasoning benchmark focused on deductive and causal inference, testing a model's ability to apply rule-based logic to natural language contexts. We use the official `test` split from the LogiQA2.0 release on GitHub.
- **TruthfulQA** [21]: a benchmark evaluating factual accuracy against common misconceptions and hallucinations. We reformulate both MC1 and MC2 variants into a unified multiple-choice format, synchronously shuffling options and labels for each sample to ensure consistent randomization across runs.
- **FACTOR** [25]: a factual consistency benchmark comprising *wiki*, *news*, and *expert* splits, designed to measure whether model-generated statements align with known facts and avoid internal contradictions. For each example, we place the correct answer first and concatenate it with up to three contradicted statements to form a multiple-choice question. The options are then deterministically shuffled to ensure reproducible evaluation.

A.10 Conditional Accuracy Analysis

Table 11 details the trade-off between coverage and precision. The Non-Expert baseline is highly precise but overly conservative. RSN steering relaxes this threshold to match the Expert profile.

Table 11: Micro-Dynamics of Confidence (LLaMA3-8B-IT). RSN injection reduces the Non-Expert's excessive conditional accuracy to a decisive level.

Config	Total Acc	E-ratio	Cond. Acc
Non-Exp.	38.65	44.77	**69.33**
$+4RSN_{mid}$	**63.22**	**3.73**	65.45
Expert Ref.	*63.01*	*6.87*	*67.22*

A.11 Cross-Model Consistency of Confidence Enhancement

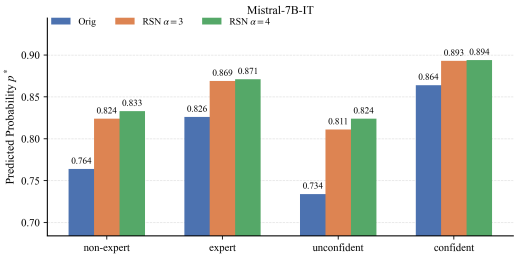
To verify that the confidence-enhancing effect of RSN editing is not specific to a single architecture,

we replicate the probability analysis from Figure 5 on *Mistral-7B-IT* and *Qwen3-8B-IT*.

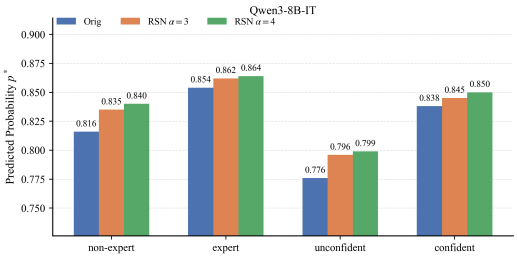
Figure 11 reports the mean predicted probability (p^*) across four prompting conditions (*non-expert*, *expert*, *unconfident*, *confident*) before and after RSN intervention ($\alpha = 3, 4$).

Across both models, RSN editing consistently increases p^* relative to the unedited baseline, mirroring the monotonic trend observed in LLaMA3 (Figure 5). The largest gains again appear in *non-expert* and *unconfident* settings, confirming that RSNs strengthen internal certainty where the baseline confidence is weakest. Meanwhile, already confident conditions show smaller yet stable improvements, suggesting that RSN scaling saturates once activation approaches a high-certainty regime.

These results demonstrate that the observed confidence amplification is architecture-agnostic: RSN editing provides a consistent, model-general mechanism for increasing internal decisiveness without relying on linguistic cues or fine-tuning.



(a) *Mistral-7B-IT*



(b) *Qwen3-8B-IT*

Figure 11: Cross-Model Confidence Enhancement. Predicted probability (p^*) under RSN editing. RSN intervention ($\alpha = 3, 4$) consistently increases internal confidence across all roles and prompting conditions, replicating the trend shown for LLaMA3 in Figure 5.

A.12 Verbalized Confidence and Entropy Analysis on MMLU-Pro

To validate whether the shift in internal states (MSP) translates to explicit output, we conduct a comprehensive analysis on *Standard MMLU-Pro* [40]. We focus on two complementary metrics: the distributional shift of explicit confidence scores, and the entropy (stability) of these predictions.

Distributional Shift under RSN. We first evaluate the model's "Reasoning Willingness" (self-evaluation on a 0–9 scale) under the *Neutral* condition. As shown in Figure 12, the original model exhibits a broad and highly variable distribution (Mean $\approx 4.9 \pm 1.91$), reflecting unstable certainty when explicit cues are absent. Crucially, RSN injection ($\alpha = 3, 4$) transforms this into a compact, high-confidence regime (Mean $\approx 7.6 \pm 0.55$). This contraction of variance and upward shift confirms that RSNs fundamentally restructure the model's internal state to be more decisive.

Entropy and Stability. To verify that this increased confidence reflects stable internal states rather than random noise, we analyze the entropy of the self-generated scores. Figure 13 visualizes the relationship between mean scores and entropy across five prompting conditions. A distinct inverse correlation is observed: high-confidence roles (e.g., *confident*, *expert*) occupy low-entropy regions, while low-confidence settings exhibit higher entropy. This confirms that the "confidence" steering not only shifts the mean score but also sharpens the underlying activation distribution.

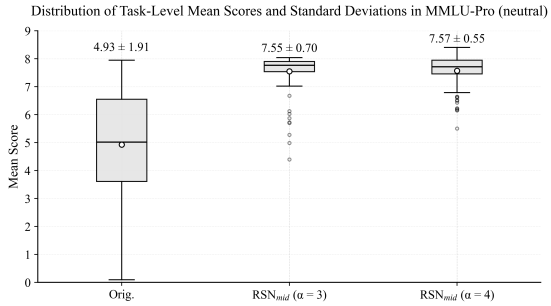


Figure 12: RSN Effects on Self-Evaluation (Neutral Condition). On *MMLU-Pro* (90 tasks), positive RSN shifts (α=3, 4) raise and compress self-evaluated scores, indicating stabilized internal activations.

A.13 Extended Analysis on Mistral-7B-IT

In the main text, we noted that Mistral-7B-IT exhibits distinct dynamics compared to LLaMA3 and Qwen3. Here, we provide a comprehensive analysis of this boundary condition, combining performance metrics with mechanistic entropy analysis.

A.13.1 Performance Dynamics: The Factuality Trade-off

As detailed in Table 12, positive RSN steering ($\alpha = 4$) on Mistral does not yield gains on reasoning tasks and significantly impairs factuality. This suggests that blind confidence boosting is ineffective when the base model lacks sufficient latent representations. However, consistent with LLaMA3, negative steering ($\alpha = -4$) effectively functions as a

Table 12: Mistral-7B-IT Performance in Neutral Settings. Positive steering ($\alpha = 4$) fails to improve reasoning, whereas negative steering ($\alpha = -4$) recovers factual accuracy. This highlights that RSNs modulate decision threshold rather than knowledge.

Condition	**Reasoning Benchmarks (Acc %)**					**Factuality (Acc %)**		
	MMLU	MMLUPro	GPQA	AR-LSAT	LogiQA	TQA-MC1	TQA-MC2	FACTOR
Orig	59.42	**31.67**	30.34	21.47	50.00	46.27	57.65	66.98
+RSN ($\alpha = 4$)	58.03	28.25	**30.65**	21.23	49.81	45.90	57.04	61.97
+RSN ($\alpha = -4$)	–	30.10	29.26	20.80	**51.15**	45.90	**59.73**	**68.04**

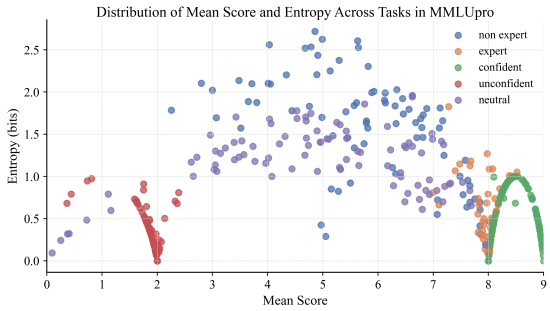


Figure 13: Entropy vs. Self-Evaluation Score (MMLU-Pro). Task-level mean scores and entropy values across five prompting conditions. Higher self-confidence systematically correlates with lower entropy (higher stability), confirming that RSN-induced confidence represents a stable, low-entropy state.

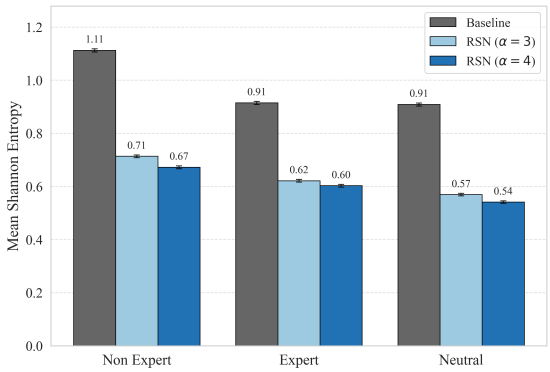


Figure 14: Decoupling Confidence from Capability (Mistral-7B-IT). While accuracy remains flat (Table 12), RSN editing triggers a drastic drop in Constrained Entropy. This confirms that RSNs amplify certainty independent of correctness.

"caution filter," improving performance on Factuality benchmarks (TruthfulQA-MC2 and FACTOR).

A.13.2 Mechanism Isolation: The Case of "Unwarranted Certainty"

To understand *why* accuracy remains flat despite steering, we analyze the model's internal uncertainty distribution via *Constrained Entropy* on MMLU-Pro.

As shown in Figure 14, RSN injection triggers a precipitous drop in entropy across all conditions, even though accuracy (Table 12) does not improve. This decoupling confirms a critical mechanistic insight: RSNs function purely as a *gain control mechanism*. They sharpen the probability distribution to force a decision, but they cannot inject new knowledge. When the underlying knowledge is absent, RSNs simply create "unwarranted certainty," confirming that the mechanism modulates *willingness* rather than *capability*.

Code and scripts will be released at anonymized URL upon acceptance.