Where Does Robustness Live? Neuron-Guided Adaptation for Retrieval-Augmented Language Models

Anonymous ACL submission

Abstract

Retrieval-Augmented Language Models (RALMs) have shown strong potential in knowledge-intensive tasks, yet they remain vulnerable when retrieved contexts are noisy or irrelevant. Robustness against such contexts requires two distinct capabilities: *abstention* when contexts are uninformative, and *selective extraction* when relevant evidence is buried in noise. Yet existing methods face two key limitations: they do not train separately for these two capabilities, and they adapt the model at a coarse layer- or module-level granularity, overlooking that only a small subset of neurons is strongly activated for a given input. We propose NeuRIT, a **Neu**ron-guided **R**obust **I**nstruction-**T**uning framework built on a *localization-first* perspective. NeuRIT mines context-aware neurons associated with relevant and irrelevant context processing, and uses them as *anchors* to selectively adapt both the identified neuron groups and the layers in which they concentrate. NeuRIT then performs two-stage instruction tuning that teaches complementary behaviors: suppress generation when there is nothing to extract, and extract relevant evidence when there is. NeuRIT consistently outperforms strong baselines while updating only around 6.6% of model parameters.

1 Introduction

While Large Language Models (LLMs) have achieved remarkable success, they still face significant limitations, including their tendency to produce hallucinations in knowledge-intensive tasks (*e.g.,* question-answering) (Wang and Yu, 2025; Huang et al., 2025). To mitigate these problems, Retrieval-Augmented Language Models (RALMs) (Lewis et al., 2020) provide a promising solution. RALMs operate in two main steps: (1) invoking a retriever to search for and extract relevant contexts to a given query from external sources, and (2) leveraging the retrieved contexts to prompt LLMs, thereby generating more accurate and grounded responses (Guu et al., 2020; Izacard et al., 2023; Gao et al., 2023; Fan et al., 2024). Despite their ability to support grounded generation, RALMs face a significant challenge: query-irrelevant or incorrectly retrieved contexts can mislead the LLM and consequently degrade its performance (Wang et al., 2023; Cuconasu et al., 2024; Chen et al., 2024), as LLMs are not explicitly trained to be robust against such noise (Jiang et al., 2024; Lin et al., 2025).

Importantly, retrieved contexts can fail in two qualitatively distinct ways, each demanding a fundamentally different model behavior. They may be uninformative to the query, providing no usable evidence at all, in which case the model must abstain from extracting anything and rely on a refusal-like response. Alternatively, they may be partially relevant, containing the answer alongside distracting noise, in which case the model must selectively extract the useful fragment while ignoring the surrounding noise. A model trained without this distinction risks either over-suppressing useful evidence or under-suppressing noise. Robust generation thus demands two distinct capabilities: *abstention* and *selective extraction*.

Existing approaches have attempted to address this challenge in various ways. While prior work has explored refining retrieved contexts before generation (Glass et al., 2022; Vig et al., 2022; Xu et al., 2023; Wu et al., 2025b), a more direct line of work enhances the robustness of the generator LLM itself through instruction tuning (Wei et al., 2024; Lin et al., 2023; Wu et al., 2025a), teaching the model to better utilize retrieved contexts while disregarding query-irrelevant ones. However, we observe two key limitations. *First, in terms of the training objective*, existing methods do not train separately for *abstention* and *selective extraction*, leaving the tension between the two unresolved.

Second, in terms of the granularity of adaptation, they typically update parameters at the layer, module, or low-rank level (Wu et al., 2025a; Yoran et al., 2024; Xu et al., 2024), neglecting a well-established property of LLMs: only a small subset of neurons is strongly activated for a given task or input (Li et al., 2022; Xu et al., 2025). This suggests that robustness-related computations might be concentrated in limited regions of the network, and updating parameters without this consideration may unnecessarily interfere with unrelated capabilities. One possible solution is to identify robustness-related neurons and update them selectively. However, neuron-level updates alone are unlikely to reliably reshape model behavior, since model computation is inherently distributed across many parameters (Leng and Xiong, 2025). We therefore use identified neurons not as isolated update targets, but as *anchors* for adaptation: neuron-level attribution identifies context-aware behaviors underlying robustness (*i.e.,* processing relevant or irrelevant contexts), and guides updates to both these neuron groups and the layers in which they are densely concentrated.

To this end, we propose NeuRIT (**Neu**ron-guided **R**obust **I**nstruction **T**uning), an instruction-tuning framework built on a localization-first perspective that addresses both limitations above. To address the granularity limitation, NeuRIT first performs *context-aware neuron mining*, which identifies neurons (*i.e.,* existing in FFN) associated with relevant and irrelevant context processing through attribution-based analysis. Since some neurons respond to both context types, we separate them into a shared group rather than assigning them to either context-specific group, preserving the discriminative signals of relevant- and irrelevant context-aware neurons. To separate these objectives, NeuRIT then performs *neuron-guided instruction tuning*, which teaches the model two complementary behaviors—suppression and evidence extraction—through two training stages. In the suppression stage, NeuRIT updates only irrelevant context-aware neurons to suppress generation under uninformative contexts. In the extraction stage, NeuRIT further adapts all three identified neuron groups and the layers in which they concentrate, teaching the model to extract relevant evidence from partially noisy contexts. Together, these stages yield a unified robustness capability: suppress when there is nothing to extract, and extract when there is.

Our contributions are summarized as follows:

- We introduce a new perspective on robustness-oriented adaptation for RALMs: neuron attribution provides a localization signal for identifying where adaptation should occur.
- We propose NeuRIT, a two-stage neuron-guided instruction-tuning framework that uses mined context-aware neurons as *anchors* to jointly adapt the identified neuron groups and the layers in which they concentrate, while training separately for two complementary behaviors: suppression under uninformative contexts and selective extraction from partially noisy contexts.
- Extensive experiments on various QA benchmarks and different generator backbones show consistent gains over strong baselines under both in-domain and out-of-domain settings, while updating only around 6.6% of model parameters. Our analysis further reveals that neuron-guided localization drives the gains.

2 Related Work

Refining Retrieval Contexts. Prior works have proposed various context refinement methods to enhance the quality of retrieved knowledge. For example, reranking retrieved documents (Glass et al., 2022), extracting query-related content from the retrieved documents (*i.e.,* summarization) (Vig et al., 2022; Xu et al., 2023), and iterative retrieval along with intermediate reasoning paths (Trivedi et al., 2023; Wang et al., 2024; Wu et al., 2025b) have been explored. Recently, Wu et al. (2025b) proposed an advanced knowledge refinement method that incorporates re-ranking signals in generating Chain-of-Thought (CoT)-based summarization. However, many existing works have created extra modules for the knowledge refinement, which introduce additional steps beyond generation on RALM. These additional steps introduce extra computational costs, making it less practical for large-scale deployment (Wu et al., 2025a). Moreover, these additional steps may suffer from cascading errors (Yu et al., 2022), which ultimately limit their effectiveness in real-world applications.

Robustness to Retrieved Contexts in RALM. Previous studies have sought to improve the robustness of RALMs to noisy or irrelevant retrieved contexts. Yoran et al. (2024) train language models to directly generate correct answers despite input noise, implicitly enhancing robustness. More

recent works leverage instruction tuning to make the denoising process explicit: Wei et al. (2024) prompts the model to synthesize rationales that articulate the reasoning path from noisy contexts to the ground-truth answer, while Wu et al. (2025a) performs instruction tuning on data refined through a citation rewrite mechanism that uses NLI models to ensure the model cites only valid evidence. However, these approaches share two limitations. First, they collapse the qualitatively distinct failure modes of retrieval, fully irrelevant versus partially relevant contexts, into a single training objective, leaving the tension between abstention and selective extraction unresolved. Second, they rely on dense optimization strategies such as full-parameter fine-tuning or layer-level parameter-efficient fine-tuning, overlooking the neuron-level sparsity of LLMs and thus failing to selectively update the specific neurons critical for noise robustness.

3 Methodology

Problem Setup. Retrieval-Augmented Language Models (RALMs) for knowledge-intensive tasks, such as question-answering, aim to generate answers using a generator LLM, based on a query and documents returned by a retriever. Formally, let $\mathcal{D}_{\text{Q}} = \left\{(q_i, \{d_{i,j}\}_{j=1}^{k})\right\}_{i=1}^{N}$ be a set of open-domain questions, where q_i denotes a query and $\{d_{ij}\}_{j=1}^{k}$ is the set of top-k retrieved documents associated with q_i. Note that the documents are ranked based on relevance score with the given query. Then, a generator LLM $\mathcal{M}$ is expected to produce the output $\hat{y}$ as follows: $\hat{y}_i = \mathcal{M}(\texttt{Prompt}_i)$ where $\texttt{Prompt}_i = (q, \{d_{i,j}\}_{j=1}^{k})$ refers to a predefined prompt template along with inputs. However, the retrieved documents set $\{d_{i,j}\}_{j=1}^{k}$ often contain query-irrelevant documents that lead to performance degradation. To this end, we propose to identify relevant and irrelevant context-aware neurons and enhance their robustness to retrieved information. Our method consists of two primary components: 1) context-aware neuron mining, and 2) robustness enhancement instruction tuning on the identified neurons. The overall architecture is illustrated in Figure 1.

3.1 Context-Aware Neuron Mining

Attribution-Based Context-Aware Neuron Identification. To decide where robustness-oriented adaptation should occur, we first quantify which FFN neurons contribute to processing relevant and irrelevant retrieved contexts. To quantify the contribution of each neuron to processing relevant and irrelevant contexts, we leverage the attribution method of Shi et al. (2024). For each query q_i, we first construct a triplet $(q_i, d_i^{\text{rel}}, d_i^{\text{irrel}})$. Here, d_i^{rel} denotes a top-ranked relevant document, and d_i^{irrel} denotes an irrelevant document with a zero relevance score. Next, for every query q_i and document d_i^c with context type $c \in \{\text{rel,irrel}\}$, we measure the attribution score of every neuron n (that exists in FFNs) to the prediction y. This is done via Integrated Gradients (Sundararajan et al., 2017). Specifically, let $v(q_i)$ denote the activation of neuron n given only the query, and $v(q_i, d_i^c)$ denote the activation given both the query and the document of context type c. We compute the attribution score as follows:

$$\mathcal{A}(n; q_i, d_i^c) = \big(v(q_i, d_i^c) - v(q_i)\big) \times \int_{\alpha=0}^{1} \frac{\partial P(y \mid q_i, d_i^c, v_\alpha)}{\partial v_\alpha}\, d\alpha, \quad (1)$$

where $v_\alpha = v(q_i) + \alpha\big(v(q_i, d_i^c) - v(q_i)\big)$ interpolates between the two activation states for $\alpha \in [0, 1]$, and the integral is approximated using a 20-step Riemann sum. Based on these scores, for each query, we filter for the top 10% and select the top 20 highest-scoring neurons. We then aggregate these neuron selections across all queries within each context type, and identify the top-t most frequently occurring neurons. We denote these initial candidate sets of neurons as $\tilde{\mathcal{P}}_{\text{irrel}}$ and $\tilde{\mathcal{P}}_{\text{rel}}$ for irrelevant and relevant contexts, respectively (See Figure 1, left).

Decoupling Context-Aware Neurons. Once we obtain $\tilde{\mathcal{P}}_{\text{rel}}$ and $\tilde{\mathcal{P}}_{\text{irrel}}$, we explore the intersection between the two sets. We observe that identified neuron sets $\tilde{\mathcal{P}}_{\text{irrel}}$ and $\tilde{\mathcal{P}}_{\text{rel}}$ reveal a significant overlap[1]. This is natural since parameters associated to neurons in LLMs form distributed and overlapping representations rather than being functionally exclusive (Leng and Xiong, 2025). However, such overlapping neurons do not provide discriminative signals for identifying context-specific awareness, as their activations cannot be uniquely attributed to either relevant or irrelevant contexts. To address this, we define distinct neuron sets by excluding

[1]Initial mining yields 7,380 relevant and 7,036 irrelevant context-aware candidate neurons with a significant overlap of 6,240 neurons.

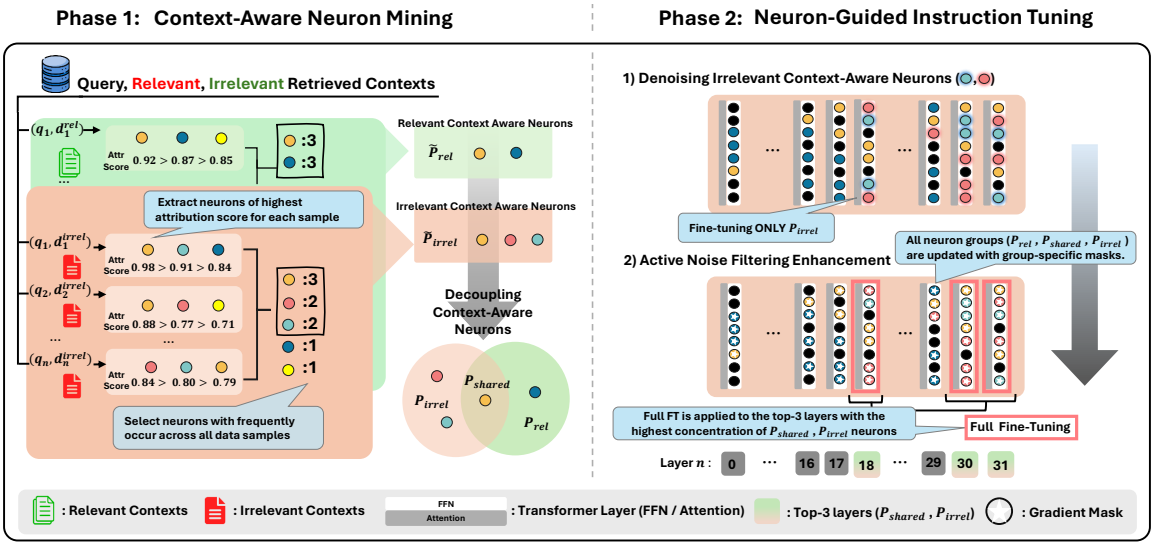

Figure 1: Overview of the NeuRIT framework. NeuRIT consists of two phases: (1) Context-Aware Neuron Mining, where neurons are categorized into relevant, irrelevant, and shared groups based on attribution scores from retrieved contexts; and (2) Neuron-Guided Instruction Tuning, where the identified neuron groups guide the training process through group-specific updates.

neurons that exhibit high attribution in both context types. Specifically, we define the overlapping neurons set as $\mathcal{P}_{\text{shared}} = \tilde{\mathcal{P}}_{\text{rel}} \cap \tilde{\mathcal{P}}_{\text{irrel}}$, and remove them from each set as follows: $\mathcal{P}_{\text{rel}} \leftarrow \tilde{\mathcal{P}}_{\text{rel}} \setminus \mathcal{P}_{\text{shared}}$, $\mathcal{P}_{\text{irrel}} \leftarrow \tilde{\mathcal{P}}_{\text{irrel}} \setminus \mathcal{P}_{\text{shared}}$. This procedure ensures that the resulting neuron sets capture distinct context-specific awareness for relevant and irrelevant contexts, respectively (See Figure 1, left). Our mining results in 100 neurons for $\mathcal{P}_{\text{rel}}$ and $\mathcal{P}_{\text{irrel}}$ and 30 neurons for $\mathcal{P}_{\text{shared}}$.

3.2 Neuron-Guided Instruction Tuning

Based on the identified neurons, we propose a two-stage neuron-guided instruction tuning framework, with each stage targeting one robustness behavior: (1) a denoising stage that teaches abstention by suppressing irrelevant context-aware neurons under uninformative contexts, and (2) a noise filtering stage that teaches selective extraction by adapting the identified neuron groups within the layers where noise-related neurons are most densely concentrated.

Denoising Irrelevant Context-Aware Neurons. For uninformative retrieved contexts, robust generation requires abstention rather than producing answers from unsupported evidence. To induce this behavior at the neuron level, we instruction-tune the irrelevant context-aware neurons $\mathcal{P}_{\text{irrel}}$ to emit an End-of-Text (EOT) token under irrelevant-context inputs. Since $\mathcal{P}_{\text{irrel}}$ are identified as primarily responsible for processing irrelevant or misleading contexts, allowing them to participate in processing relevant information poses a risk of latent interference. Accordingly, we impose a hard constraint that decouples these irrelevant context-aware neurons from generation.

Noise Filtering Enhancement on Neuron-level Parameters. In partially relevant retrieval settings, abstention alone is insufficient because useful evidence may be mixed with distracting noise. We therefore adapt the model where noise-robustness computations are concentrated. This is because neuron-level updates alone are unlikely to reliably reshape model behavior, since model computation is inherently distributed across many parameters (Leng and Xiong, 2025). Hence, we update the top-3 layers with the highest density of $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$, as these layers aggregate noise-discriminative signals near the output[2]. Within these localized layers, we further update the identified neuron groups ($\mathcal{P}_{\text{rel}}$, $\mathcal{P}_{\text{shared}}$, $\mathcal{P}_{\text{irrel}}$), allowing the model to refine relevant-evidence extraction while controlling interference from irrelevant contexts. Specifically, we optimize the selected parameters to make the generator produce a distilled summary that contains only query-relevant information from potentially noisy retrieved contexts. Given an original query and its retrieved

[2]See § 4.5 for the layer-wise distribution analysis motivating this choice.

contexts, the model is instructed to filter out irrelevant content and generate only the relevant evidence needed to answer the query. This aligns the localized adaptation with our context-filtering objective. For this, we construct a dataset as follows: for a pair $(q_i, \{d_{i,j}\}_{j=1}^{k}) \in \mathcal{D}_Q$, we employ an external LLM[3] to generate Relevant Summaries (RS) y_i^{RS}, which includes only query-relevant information from $\{d_{ij}\}_{j=1}^{k}$ by giving a specific instruction[5]. We construct an instruction-tuning dataset as follows:

$$\mathcal{D}_{\text{RS}} = \left\{ \left(q_i, \{d_{ij}\}_{j=1}^{k}),\ y_i^{\text{RS}} \right) \right\}_{i=1}^{n}$$

Once done training, to enable open-domain question answering, we design a dual-instruction prompt consisting of (1) relevant summary extraction ($\texttt{inst}_{\text{RSE}}$: `extract relevant information from provided documents`) and (2) a question answering instruction ($\texttt{inst}_{\text{QA}}$: `Answer a question as briefly as possible`). Each training input x_i is formatted as a prompt combining these components:

$$\texttt{Prompt}_i = (\texttt{inst}_{\text{RSE}}, \texttt{inst}_{\text{QA}}, \{d_{ij}\}_{j=1}^{k}, q_i)$$

The instruction-tuned generator LLM produces the corresponding answer as $\hat{y}_i = \mathcal{M}(\texttt{Prompt}_i)$. This design jointly enables the model to filter noisy documents, extract and align relevant evidence, and generate factually grounded answers.

4 Experiments

4.1 Data

We evaluate our method on a diverse QA benchmarks covering various question types and reasoning settings. We use KILT-formatted versions of Natural Questions, TriviaQA, and HotpotQA (Rau et al., 2024a). **KILT-NQ (Natural Questions)** (Kwiatkowski et al., 2019) contains real anonymized Google search queries with answer annotations on Wikipedia pages. **ASQA** (Stelmakh et al., 2022) is a long-form QA benchmark for ambiguous factoid questions, requiring answers to synthesize evidence from multiple sources and resolve different interpretations. **KILT-TriviaQA** (Joshi et al., 2017) is a factoid QA benchmark derived from TriviaQA, with answer provenance aligned to Wikipedia evidence. **SciQ** (Welbl et al., 2017) is a multiple-choice science QA dataset consisting of crowdsourced exam-style questions with supporting evidence paragraphs. **POPQA** (Mallen et al., 2023) evaluates factual knowledge about entities with varying popularity, especially long-tail entities. **KILT-HotpotQA** (Yang et al., 2018) requires multi-hop reasoning over multiple Wikipedia pages. **2WikiMultiHopQA** (Ho et al., 2020) requires connecting information across multiple Wikipedia articles to answer entity-centric multi-hop questions.

4.2 Baselines

To validate the effectiveness of NeuRIT, we conduct a comparative evaluation against three distinct categories of baselines: 1) Standard RALM, 2) Robustness Enhancing Tuning for RALMs, and 3) Retrieved Context Refinement. For the refinement methods, we additionally evaluate them in a compositional manner by attaching each refinement module to every baseline in the first two categories.

Standard RALM. This baseline follows the conventional Retrieval Augmented Language Model (RALM) generation paradigm, where retrieved documents are concatenated along with a given query without any additional control mechanisms.

Robustness Enhancing Tuning for RALMs. These methods tune the generator LLM directly to be robust to irrelevant or misleading retrieved contexts: **RetRobust**[6] (Yoran et al., 2024) trains the generator LLM on noisy retrieval inputs to directly produce correct answers despite distracting or misleading contexts, encouraging robustness to irrelevant evidence without explicit denoising supervision. **InstructRAG**[7] (Wei et al., 2024) instead applies instruction tuning on self-synthesized rationales that explain how ground-truth answers are derived from noisy retrieved documents, providing explicit denoising supervision. **PA-RAG**[8] (Wu et al., 2025a) combines instruction fine-tuning with multi-stage DPO that sequentially optimizes response informativeness, response robustness, and citation quality.

Retrieved Contexts Refinement. In contrast to methods that directly tune the generator LLM for robustness, these baselines improve robustness by refining retrieved contexts before they are given to

[3] In our experiments, we used `GPT-4.1.mini` (OpenAI 2025)[4] though any LLM can be used.

[5] See Table B.1 in Appendix for details

[6] https://huggingface.co/Ori/llama-2-13b-peft-nq-retrobust

[7] https://huggingface.co/meng-lab/2WikiMultiHopQA-InstructRAG-FT

[8] https://huggingface.co/wuqiong1/PA-RAG_Meta-Llama-3-8B-Instruct

Method	NQ	ASQA	TriviaQA	SCIQ	POPQA	HotpotQA	2Wiki	Avg.
RALM (Lewis et al., 2020)	62.21	67.82	88.60	53.30	60.85	45.44	42.27	60.07
+ Reranker (Glass et al., 2022)	61.65	67.93	88.31	52.60	60.37	45.27	41.84	59.71
+ RankCoT (Wu et al., 2025b)	59.43	65.93	89.60	57.40	59.80	48.54	43.42	60.59
RetRobust (Yoran et al., 2024)	53.23	61.92	89.80	58.20	53.24	40.84	35.17	56.06
+ Reranker (Glass et al., 2022)	54.42	64.13	90.55	58.20	54.92	41.46	35.69	57.05
+ RankCoT (Wu et al., 2025b)	53.26	60.55	88.93	57.80	53.18	43.84	39.94	56.79
InstructRAG (Wei et al., 2024)	60.62	66.98	87.48	55.30	60.25	45.33	44.61	60.08
+ Reranker (Glass et al., 2022)	63.58	71.51	90.73	56.80	64.60	48.32	47.01	63.22
+ RankCoT (Wu et al., 2025b)	60.17	67.72	90.76	59.50	62.24	51.66	50.06	63.16
PA-RAG (Wu et al., 2025a)	67.92	74.47	89.86	56.60	64.37	49.29	46.57	64.15
+ Reranker (Glass et al., 2022)	72.01	79.43	92.58	57.60	71.08	54.12	48.87	67.96
+ RankCoT (Wu et al., 2025b)	66.30	73.10	92.32	**63.40**	65.68	53.89	**52.02**	66.67
NeuRIT	69.68	75.31	92.21	60.80	67.37	52.25	46.34	66.28
+ Reranker (Glass et al., 2022)	**72.57**	**80.48**	**93.91**†	61.60	**74.29**†	**55.71**†	47.80	**69.48**
+ RankCoT (Wu et al., 2025b)	61.93	69.09	91.60	59.10	63.28	52.04	51.13	64.02

Table 1: Main results of performance comparison across QA benchmarks. We report **Accuracy** for each dataset. †: NeuRIT significantly improves the best baseline at $p < 0.05$ with paired t-test. The best results are highlighted in **bold** and the second-best results are underlined (per metric).

the generator LLM. **Reranker**[9] (Glass et al., 2022) uses a cross-encoder re-ranker to prioritize query-relevant documents, thereby reducing the influence of irrelevant contexts in the prompt. **RankCoT**[10] (Wu et al., 2025b) instead refines retrieved documents by combining re-ranking signals with CoT-based summarization, serving as a stronger noise-mitigation baseline at the cost of additional reasoning overhead.

4.3 Implementation Details

All experiments are conducted using the generator LLM as LLaMA-3-8B-Instruct (Dubey et al., 2024), and training is performed on a single NVIDIA H200 GPU with 141GB memory. To identify context-aware neurons, we use approximately 400 sub-samples of HotpotQA training data. We set frequency occurrence threshold k as 130 to ensure 100 neurons are selected for both $\mathcal{P}_{\text{rel}}$ and $\mathcal{P}_{\text{irrel}}$ and 30 neurons for $\mathcal{P}_{\text{shared}}$. For $\mathcal{D}_{\text{RS}}$, the external LLM produces 142-token summaries grounded exclusively in the provided documents, with no intrinsic knowledge use enforced via prompting to mitigate hallucination. Detailed training, implementation, and neuron selection hyperparameter settings are reported in Appendix B.2 and B.3.

4.4 Main Results

We report results in Table 1 for three variants: the base model, the base model with Reranker, and the base model with RankCoT. Without ranking, NeuRIT consistently outperforms most baselines and remains comparable to or slightly better than the strongest one, PA-RAG, across all datasets. Notably, this strong base performance is observed despite NeuRIT being trained exclusively on HotpotQA, indicating that the learned behavior generalizes beyond the training distribution to the remaining six datasets. When the base model is augmented with a Reranker, performance is consistently improved for RetRobust, InstructRAG, PA-RAG, and NeuRIT across all benchmarks, with particularly pronounced gains on single-hop datasets such as NQ, ASQA, and TriviaQA. Under this setting, NeuRIT achieves the highest average accuracy of 69.48, outperforming all baseline configurations. In contrast, augmenting the base model with reasoning-aware ranking via RankCoT yields a different performance pattern: while its impact is limited on single-hop benchmarks, it provides the greatest improvements on 2Wiki, which requires multi-hop reasoning over longer contexts. We further evaluate our method using LLMEval (Rau et al., 2024a) (Appendix C.2), which confirms that NeuRIT consistently achieves superior answer accuracy and evidence faithfulness.

4.5 Analysis

Efficiency Analysis. NeuRIT is efficient in two respects. In terms of parameter efficiency, it updates only 0.529B of LLaMA-3-8B-Instruct's 8B parameters ($\sim$6.6%). In terms of compute and cost, the entire pipeline (*i.e.,* neuron mining, summary

[9] https://huggingface.co/naver/trecdl22-crossencoder-debertav3

[10] https://huggingface.co/MignonMiyoung/RankCoT

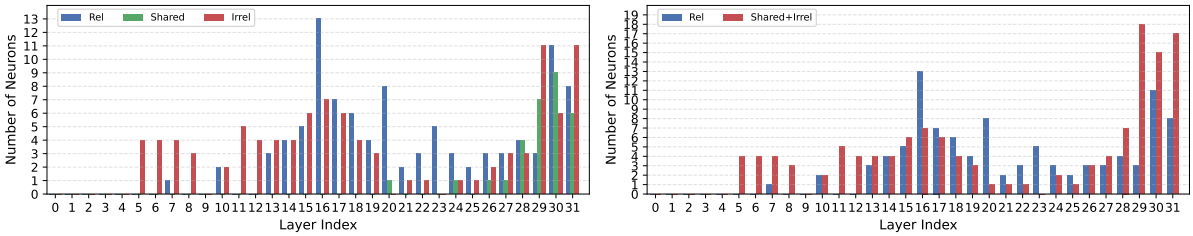

Figure 2: Neuron distribution of LLaMA-3-8B-Instruct on the HotpotQA dataset; the left panel shows the distribution of $\mathcal{P}_{\text{rel}}$, $\mathcal{P}_{\text{shared}}$ and $\mathcal{P}_{\text{irrel}}$, while the right panel shows the $\mathcal{P}_{\text{rel}}$ against the combination of $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$.

Mistral-7B	NQ	ASQA	TriviaQA	SCIQ	POPQA	HotpotQA	2Wiki	AVG.
RALM	64.85	71.20	89.59	60.80	62.78	48.39	48.80	63.77
NeuRIT	**66.47**	**72.57**	**90.04**	**62.50**	**64.15**	**50.46**	**50.65**	**65.26**

Qwen-2.5-14B	NQ	ASQA	TriviaQA	SCIQ	POPQA	HotpotQA	2Wiki	AVG.
RALM	66.33	72.99	91.29	59.40	63.72	52.73	**50.97**	65.35
NeuRIT	**70.42**	**75.94**	**92.49**	**67.00**	**66.36**	**53.42**	50.32	**67.99**

LLaMA-2-13B	NQ	ASQA	TriviaQA	SCIQ	POPQA	HotpotQA	2Wiki	AVG.
RALM	60.84	67.51	88.27	57.40	55.35	45.66	44.42	59.92
RetRobust	53.23	61.92	89.80	58.20	53.24	40.84	35.17	56.06
NeuRIT	**69.15**	**75.31**	**90.91**	**63.20**	**65.31**	**50.91**	**44.87**	**65.66**

Table 2: Comparison of accuracy across different backbone models. (Top) Mistral-7B-Instruct-v0.2. (Middle) Qwen-2.5-14B-Instruct. (Bottom) LLaMA-2-13B. Best results are in **bold**.

	Relevant			Irrelevant		
Method	**NQ**	**HotpotQA**	**SCIQ**	**NQ**	**HotpotQA**	**SCIQ**
RALM	79.16	73.15	71.08	20.43	15.19	17.49
InstructRAG	76.87	74.29	75.04	19.54	11.98	16.62
PA-RAG	86.38	**81.96**	79.45	20.30	12.32	12.83
NeuRIT	**86.73**	81.75	**82.50**	**25.76**	**18.45**	**20.41**

Table 3: The comparison of the accuracy of our method in scenarios with relevant (answer-present) and irrelevant (answer-absent) retrieved contexts

generation, and two-stage instruction tuning) completes in 17 hours on a single H200 GPU, with the `GPT-4.1.mini` summary dataset costing only $2.81 to construct. This combination of selective adaptation and low overhead makes NeuRIT practical to deploy.

Layer-wise Distribution of Identified Neurons. We select the top-3 layers with the highest density of $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$ as the noise filtering targets. This choice is based on the following observations from Figure 2. First, $\mathcal{P}_{\text{rel}}$ is mainly concentrated in the middle layers, peaking around Layers 15–19. In contrast, $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$ are most prominent in the upper layers, especially Layers 29–31, while $\mathcal{P}_{\text{irrel}}$ also spans Layers 6–18. Figure 2 (Right) further highlights this layer-wise separation by comparing $\mathcal{P}_{\text{rel}}$ with the combined distribution of $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$. These patterns suggest that irrelevant and mixed-context processing is concentrated in a small set of upper layers, motivating our top-3 density-based layer selection rather than arbitrary layer tuning.

Generalization across Generator Backbones. To test generalization across generators, we apply our method to Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Qwen-2.5-14B-Instruct (Qwen et al., 2025), and LLaMA-2-13B-Chat (Touvron et al., 2023), varying architecture and scale while keeping other components fixed. As shown in Table 2, NeuRIT consistently outperforms the RALM baseline in average accuracy over seven benchmarks across all three backbone models. We further analyze the layer-wise neuron distributions used for Neuron-Guided Instruction Tuning, with details provided in Appendix D.1. The consistent performance gains and aligned distributional patterns suggest that NeuRIT transfers across different generators and may capture a general mechanism for improving robustness in RALMs.

Irrelevant Contexts Handling. To verify the robustness of NeuRIT, we evaluate its performance in two distinct scenarios: 1) where retrieved contexts contain the ground-truth answer (Relevant), and 2) where they do not (Irrelevant). As summarized in Table 3, NeuRIT consistently outperforms all baseline methods across both settings. Notably, in the Irrelevant scenario—where model robustness is most severely tested—NeuRIT achieves the highest accuracy on all three datasets (NQ, HotpotQA, and SCIQ), significantly surpassing PA-RAG. This suggests that our approach more effectively filters out retrieval noise and mitigates the risk of hallucination without the need for additional external modules.

4.6 Ablation Study

In this section, we revisit three claims underlying NeuRIT: (i) whether each training stage is necessary; (ii) the gain stems from *neuron-guided* layer selection rather than arbitrary layer tuning; and

Method	NQ	HotpotQA	TriviaQA	SCIQ	POPQA	Avg.
NeuRIT	69.68	52.25	92.21	60.80	67.37	**68.46**
Neuron-guided layer selection vs. random selection						
Random layer sel.	65.31	45.53	89.54	55.60	62.66	63.73
Stage-level ablations						
w/o denoising	67.88	50.37	91.16	56.60	65.87	66.38
w/o NoiseFilter	61.50	44.82	87.88	52.10	59.60	61.18
w/o denoising, NoiseFilter	62.81	48.28	88.68	52.40	59.50	62.33
Component ablations within NoiseFilter						
w/o NoiseFilter, Layers	61.93	47.10	88.98	51.90	58.94	61.77
w/o NoiseFilter, Neurons	69.01	53.26	91.85	59.90	67.00	68.20
w/o NoiseFilter, $\mathcal{D}_{\text{RS}}$	67.53	51.60	91.14	58.10	65.82	66.84
Sensitivity to neuron budget						
Half $(50, 15, 50)$	70.28	50.87	92.06	60.40	67.16	68.15
Double $(200, 60, 200)$	70.60	50.96	91.61	60.80	67.37	68.27

Table 4: Ablation results of NeuRIT across five QA datasets. Average accuracy is computed over the five datasets.

(iii) within the selected layers, neuron-level masks provide a complementary refinement on top of the layer-level signal.

Stage-level contributions. We first examine whether each of the two training stages independently contributes to the final performance. As shown in Table 4, removing either the denoising stage (*w/o denoising*) or the noise filtering stage (*w/o NoiseFilter*) consistently degrades performance, dropping the average accuracy from 68.46 to 66.38 and 61.18, respectively. Notably, applying the denoising stage *alone* (*w/o NoiseFilter*: 61.18) falls below the five-dataset average of the RALM baseline (62.08), confirming that the EOT-based suppression signal can over-suppress useful evidence when partially relevant contexts are present. The subsequent noise filtering stage compensates by teaching the model to extract evidence from mixed-quality contexts, and the two stages thus operate as a composed mechanism rather than independent contributions. Furthermore, replacing this composed pipeline with standard LoRA fine-tuning on the same $\mathcal{D}_{\text{RS}}$ data (*w/o denoising, NoiseFilter*: 62.33) still falls 6.13 points behind NeuRIT, isolating the contribution of the neuron-guided framework from the distilled supervision itself.

Neuron-guided layer selection drives the gain. A natural concern is whether NeuRIT's gains can be attributed to neuron mining at all, or whether simply fine-tuning a few upper layers would suffice. To isolate this, we sample neurons uniformly at random while matching the original group sizes $(|\mathcal{P}_{\text{rel}}|, |\mathcal{P}_{\text{shared}}|, |\mathcal{P}_{\text{irrel}}|) = (100, 30, 100)$, then select the top-3 layers with the highest density of these randomly sampled neurons and apply identical two-stage tuning. As reported in Table 4, this *Random layer selection* baseline reaches only 63.73 average accuracy, a 4.73-point gap from NeuRIT (68.46) and below even the standalone PA-RAG baseline. This confirms that the primary role of neuron mining is to provide a localization signal: it identifies where in the network adaptation should occur. Without this guidance, layer-level tuning becomes arbitrary and substantially less effective.

Layer-level vs. neuron-level contributions within NoiseFilter. Within the noise filtering stage, we update two complementary categories: (i) the top-3 layers selected by neuron density, and (ii) neuron-level gradient masks applied to $\mathcal{P}_{\text{rel}}$, $\mathcal{P}_{\text{shared}}$, and $\mathcal{P}_{\text{irrel}}$ within those layers. Removing the layer-level updates (*w/o NoiseFilter, Layers*) drops performance to 61.77, whereas removing only the neuron-level masks (*w/o NoiseFilter, Neurons*) yields a smaller drop to 68.20. We emphasize that this small gap should *not* be interpreted as evidence that neuron-level signal is unimportant; rather, the neuron signal has already done most of its work *at the layer-selection stage*, identifying the layers where $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$ are concentrated. The neuron-level masks then provide a finer-grained refinement on top of this localization, consistent with the observation that FFN computations rely on distributed parameter interactions (Leng and Xiong, 2025) and cannot be effectively adapted via neuron updates in isolation. Replacing $\mathcal{D}_{\text{RS}}$ with raw QA pairs (*w/o NoiseFilter, $\mathcal{D}_{RS}$*: 66.84) further confirms the value of relevant-summary supervision as the training target.

Sensitivity to neuron budget. Finally, we examine sensitivity to the number of selected neurons by halving or doubling the original budget. Both settings (68.15 and 68.27) remain within 0.31 points of the original configuration (68.46), demonstrating that NeuRIT is robust to the precise neuron count and that the chosen budget is not finely tuned.

5 Conclusion

In this paper, we introduced NeuRIT, a neuron-guided instruction-tuning framework for robust RALMs. NeuRIT uses attribution-mined context-aware neurons as anchors to selectively adapt relevant neuron groups and their concentrated layers. By separately learning abstention and selective extraction, NeuRIT improves robustness to retrieval noise while updating only a small fraction of parameters.

Limitations

Although NeuRIT demonstrates strong robustness across diverse QA benchmarks and multiple backbone models, several limitations remain. First, while our experiments cover mid-sized open-source LLMs up to the 14B scale, the effectiveness of neuron-guided instruction tuning on substantially larger models, such as 30B or 70B-scale LLMs, remains unexplored. Since neuron-level attribution patterns may vary with model capacity and architecture, future work should examine whether the identified context-aware neurons remain stable and effective at larger scales. Second, our relevant-summary supervision relies on an external LLM, which may introduce teacher-model bias, additional cost, and sensitivity to the quality of generated summaries. Future work could investigate weaker supervision, self-training, or human-verified evidence summaries to reduce this dependency.

References

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

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

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *Preprint*, arXiv:1907.09190.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

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

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. 2024. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers. *Advances in Neural Information Processing Systems*, 37:67712–67757.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly

supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling large language models with simple yet effective depth upscaling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 23–35, Mexico City, Mexico. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. 2024. Splade-v3: New baselines for splade. *Preprint*, arXiv:2403.06789.

Yongqi Leng and Deyi Xiong. 2025. Towards understanding multi-task learning (generalization) of llms via detecting and exploring task-specific neurons. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2969–2987.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zonglin Li, Chong You, Srinadh Bhojanapalli, Daliang Li, Ankit Singh Rawat, Sashank J Reddi, Ke Ye, Felix Chern, Felix Yu, Ruiqi Guo, and 1 others. 2022. The lazy neuron phenomenon: On emergence of activation sparsity in transformers. *arXiv preprint arXiv:2210.06313*.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. Llm whisperer: An inconspicuous attack to bias llm responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024a. Bergen: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024b. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7640–7663, Miami, Florida, USA. Association for Computational Linguistics.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *Preprint*, arXiv:1803.05355.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Shuai Wang and Yinan Yu. 2025. iQUEST: An iterative question-guided framework for knowledge base question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15616–15628, Vienna, Austria. Association for Computational Linguistics.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315, Singapore. Association for Computational Linguistics.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106.

Jiayi Wu, Hengyi Cai, Lingyong Yan, Hao Sun, Xiang Li, Shuaiqiang Wang, Dawei Yin, and Ming Gao. 2025a. PA-RAG: RAG alignment via multi-perspective preference optimization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9091–9112, Albuquerque, New Mexico. Association for Computational Linguistics.

Mingyan Wu, Zhenghao Liu, Yukun Yan, Xinze Li, Shi Yu, Zheni Zeng, Yu Gu, and Ge Yu. 2025b. RankCoT: Refining knowledge for retrieval-augmented generation through ranking chain-of-thoughts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12857–12874, Vienna, Austria. Association for Computational Linguistics.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Haoyun Xu, Runzhe Zhan, Yingpeng Ma, Derek F Wong, and Lidia S Chao. 2025. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9393–9406.

Yongxin Xu, Ruizhe Zhang, Xinke Jiang, Yujie Feng, Yuzhen Xiao, Xinyu Ma, Runchuan Zhu, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Parenting: Optimizing knowledge selection of retrieval-augmented language models with parameter decoupling and tailored tuning. *arXiv preprint arXiv:2410.10360*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Neuron Attribution and Data Construction

A.1 Data for Neuron Attribution

To compute neuron attribution scores, we adopt the Integrated Gradients (IG) formulation (Sundararajan et al., 2017) and construct a dedicated binary decision dataset tailored for attribution analysis. The dataset is designed to probe neuron behaviors under contrasting retrieval conditions by explicitly controlling the relevance of the provided context.

Data Construction for Neuron Attribution We employ a standard RAG pipeline with the SPLADE-v3 (Lassance et al., 2024) retriever over the HOTPOTQA dataset (Yang et al., 2018). For each query, SPLADE-v3 assigns sparse lexical similarity scores to candidate documents. Following the triplet construction in Section 3, we construct 389 attribution instances $(q_i, d_i^{\text{rel}}, d_i^{\text{irrel}})$, where i indexes the attribution instances. The relevant context d_i^{rel} is selected as the top-ranked retrieved context for each query, while the irrelevant context d_i^{irrel} is selected from zero-scoring retrieved contexts. In cases where multiple contexts receive zero similarity scores, instances are selected according to their retrieval order. The retrieved contexts are used in their raw form without additional filtering or summarization, and all attribution instances are constructed through deterministic preprocessing using Python scripts. We then compute neuron attribution scores for relevant and irrelevant contexts separately and aggregate the selected neurons across instances, yielding the initial neuron candidate sets $\tilde{\mathcal{P}}_{\text{rel}}$ and $\tilde{\mathcal{P}}_{\text{irrel}}$.

Binary Attribution Task Formulation Each instance is formulated as a forced binary classification task. Specifically, the model is provided with a single combined input consisting of a question, a retrieved context, and a proposed answer. The model is instructed to determine whether the proposed answer can be derived solely from the given context by choosing between two predefined options: `YES` or `NO`. This formulation enforces a strict binary decision and eliminates ambiguity in the output space, enabling stable and interpretable attribution analysis. An illustrative example of the attribution dataset format is provided in Table 7.

Attribution scores are computed with respect to the loss associated with predicting the correct binary label. During IG computation, the model is restricted to inferring answer correctness based solely on the provided context, without access to any external knowledge. The resulting IG scores quantify the contribution of individual neurons to the model's binary decision and are subsequently used to identify neurons with distinct functional roles.

A.2 Examples of the Constructed Datasets

We present qualitative examples of the two instruction-tuning datasets constructed in our framework. These examples illustrate how the datasets provide complementary supervision: (i) an *irrelevant-context denoising dataset* encourages irrelevant-context suppression through EOT generation, while (ii) a *relevant summary dataset* $\mathcal{D}_{RS}$ promotes selective evidence integration by distilling only query-relevant information rather than performing generic document summarization.

Irrelevant Context Denoising Dataset. For the denoising stage, we construct training instances by selecting five retrieved documents that are irrelevant to the input query. Given these irrelevant contexts, the model is explicitly instructed to identify the absence of query-relevant information and emit an End-of-Text (EOT) token. This dataset is designed to impose a hard constraint on neurons associated with irrelevant contexts, effectively decoupling their activations from the generation process. As a result, the model learns to terminate generation when only non-informative or misleading evidence is provided. A representative example of this denoising dataset is shown in Table 8.

Relevant Summary Dataset $\mathcal{D}_{RS}$. The second dataset focuses on reinforcing the extraction of query-relevant evidence. For each query, we retrieve five candidate documents and use an external LLM to construct a concise relevant summary that integrates only the information semantically aligned with the query. For example, given the query *"Who was once considered the best kickboxer in the world, however he has been involved in a number of controversies relating to his unsportsmanlike conduct and crimes of violence outside of the ring?"*, the generated summary identifies *Badr Hari* as the target entity. Among the retrieved documents, only the passage about Badr Hari contains both the required biographical information, such as *"Moroccan-Dutch super heavyweight kickboxer"* and *"former K-1 Heavyweight champion"*, and contextual cues about unsportsmanlike conduct and violent incidents. Documents about unrelated

athletes, such as *Rau'shee Warren*, *Yamaguchi Falcão*, and *Rafael Carvalho*, are therefore excluded from the summary. An illustrative example of the constructed relevant summary is provided in Table 9.

B Training and Implementation Details

B.1 Prompt

We provide the prompts used for Denoising Irrelevant Context-Aware Neuron-level Parameters and Context-Aware Neuron-level Parameters Robustness Enforcing Fine-tuning in Section 3.2.

Prompt for Denoising Context-Aware Neuron-level Parameters

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Given a document and a query, reason step by step to identify only the parts of the document that are directly relevant to the query, and provide a concise summary of those relevant parts.

Background:
Document 1: {document 1}
Document 2: {document 2}
Document 3: {document 3}
Document 4: {document 4}
Document 5: {document 5}

Question: {question}<|eot_id|><|start_header_id|>assistant<|end_header_id|><|eot_id|>

Prompt for Context-Aware Neuron-level Parameters Robustness Enforcing Fine-tuning

<|begin_of_text|><|start_header_id|>user<|end_header_id|>

Given a document and a query, reason step by step to identify only the parts of the document that are directly relevant to the query, and provide a concise summary of those relevant parts.

Background:
Document 1: {document 1}
Document 2: {document 2}
Document 3: {document 3}
Document 4: {document 4}
Document 5: {document 5}

Question: {question}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

{relevant summary}<|eot_id|>

Prompt for Relevant Summary Data Construction

Below is a document.

Your task is to find and concisely summarize only the parts of the document that are directly relevant to the given query.

- Do not summarize the entire document.
- Exclude any information that is not related to the query.
- Focus only on the key points that are most relevant to the query.

Query: {query}

Document:
{document}

[Relevant Summary]
(For this section, reason step-by-step in a Chain-of-Thought (CoT) manner to identify the relevant information. Show your thinking process as you determine which parts of the document are relevant to the query. Then, Summarize only the information directly related to the query based on your reasoning. If nothing is relevant, leave this section blank.)

Prompt for Open-Domain Question Answering

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

You are a helpful assistant. Your task is to extract relevant information from the provided documents and answer questions as briefly as possible.<|eot_id|><|start_header_id|>user<|end_header_id|>

Background:
Document 1: {document 1}
Document 2: {document 2}
Document 3: {document 3}
Document 4: {document 4}
Document 5: {document 5}

Question: {question}<|eot_id|><|start_header_id|>assistant<|end_header_id|>

B.2 Implementation Details

All experiments are conducted using the generator LLM as LLaMA-3-8B-Instruct (Dubey et al., 2024), and training is performed on a single NVIDIA H200 GPU with 141GB memory. During data construction of the $\mathcal{D}_{\text{RS}}$, the external LLM produces 142-token summaries grounded exclusively in the provided input documents, with zero intrinsic knowledge usage enforced via prompting, thereby ensuring document-grounded generation and mitigating hallucination. Our two-stage instruction tuning both use the AdamW optimizer, with the denoising stage trained at a learning rate of 1×10^{-5} for one epoch and the noise filtering stage trained at a learning rate of 2×10^{-5} for two epochs, using a batch size of 4 throughout.

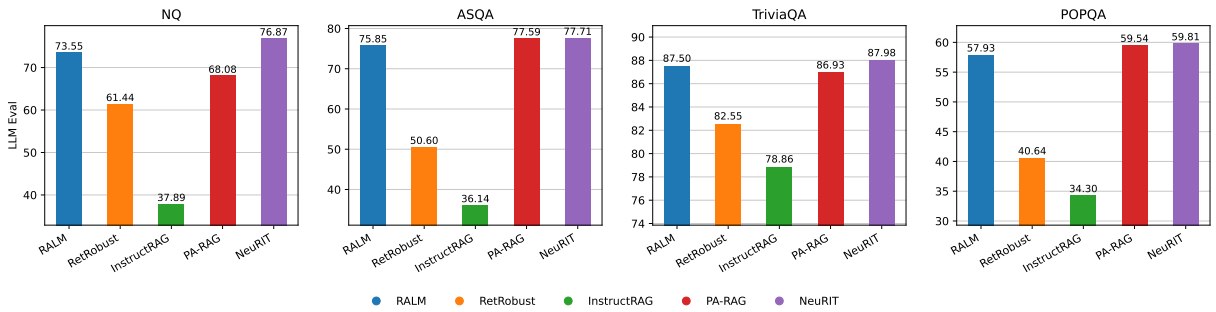

Figure 3: Performance comparison of five RAG-based methods on four QA benchmarks, evaluated using an LLM-based metric.

B.3 Neuron Mining and Selection Hyperparameters

We report the hyperparameters used for neuron mining and selection. For neuron attribution, we use 389 HOTPOTQA training instances for each context type, as described in Appendix A.1. This follows the few-hundred-sample scale used in prior attribution-based context-aware neuron analysis (Shi et al., 2024), while keeping neuron-level Integrated Gradients feasible on a single NVIDIA H200 GPU. Initial mining yields 7,380 relevant and 7,036 irrelevant candidate neurons, with 6,240 overlapping neurons; removing the overlap leaves 1,140 relevant-specific and 796 irrelevant-specific candidates. We set a compact selection budget at approximately the top 10% scale, motivated by prior representative-neuron selection practice (Shi et al., 2024). Under our frequency-based aggregation procedure, the occurrence threshold $k = 130$ yields 100 relevant and 100 irrelevant neurons, and we use the 30 most frequent overlapping neurons as $\mathcal{P}_{\text{shared}}$.

Neuron-guided Gradient Masking and Layer Adaptation. We employ stage-specific gradient masking based on the neuron groups identified by neuron mining. In the denoising stage, we target the irrelevant-context neuron set $\mathcal{P}_{\text{irrel}}$ and apply gradient masks to the FFN `up_proj`, `down_proj`, and `gate_proj` projections, thereby restricting updates to neurons associated with irrelevant-context processing. In the active noise filtering stage, we apply neuron-level gradient masks to the FFN `up_proj` and `down_proj` projections for $\mathcal{P}_{\text{rel}}$, $\mathcal{P}_{\text{irrel}}$, and $\mathcal{P}_{\text{shared}}$. Furthermore, following the analysis in Section 4.5, we fully fine-tune the top-3 layers with the highest density of $\mathcal{P}_{\text{irrel}}$ and $\mathcal{P}_{\text{shared}}$.

C Evaluation Details

C.1 Evaluation with Match Metric

All evaluations are conducted using the BERGEN framework (Rau et al., 2024a) under a retrieval-augmented generation (RAG) setting. We evaluate our method on the official test splits of the KILT-HotpotQA (Yang et al., 2018), and assess generalization across multiple benchmarks, including KILT-NQ (Kwiatkowski et al., 2019), ASQA (Stelmakh et al., 2022), KILT-TriviaQA (Joshi et al., 2017), SCIQ (Welbl et al., 2017), POPQA (Mallen et al., 2023) and 2WikiMultiHopQA (Ho et al., 2020).

No evaluation instances are used during training or neuron attribution computation. For all experiments, we fix the retriever to SPLADE-v3(Lassance et al., 2024) to ensure consistent retrieval behavior across methods. Given an input question, the retriever first retrieves the top-50 candidate documents based on sparse lexical similarity, from which the top-5 documents are provided to the generator as contextual input. To ensure fair comparison, a unified prompt template is applied across all models and datasets. Model performance is evaluated using the *Match* metric provided by the BERGEN(Rau et al., 2024b) framework, which measures exact correspondence between the generated response and the ground-truth answer. We conduct all match-based evaluations on a single NVIDIA H200 GPU with a batch size of 256 across all benchmarks.

C.2 LLM-based Evaluation Protocol

In addition to exact match evaluation, we conduct LLM-based evaluations using the BERGEN framework to assess answer quality beyond strict string matching. LLM-based evaluation is performed on four representative QA benchmarks: NQ, ASQA,

TriviaQA, and POPQA. For these datasets, we evaluate our method and baseline approaches on the development splits.

Following the BERGEN evaluation protocol, we use `SOLAR-10.7B`(Kim et al., 2024) as the evaluation model, deployed via vLLM. The LLM evaluator is configured with a batch size of 128. All evaluated methods use identical retrieved contexts and prompts to ensure a controlled comparison.

As shown in Figure 3, NeuRIT achieves the best LLM-based evaluation scores on all four benchmarks, indicating that the improvements observed under match evaluation are also reflected in answer quality judged by an LLM evaluator.

D Additional Analysis

D.1 Neuron Distribution across Generators

Figure 4 shows the layer-wise distribution of context-aware neurons across **Mistral-7B-Instruct-v0.2** (Jiang et al., 2023), **Qwen2.5-14B-Instruct** (Qwen et al., 2025), and **LLaMA-2-13B-Chat** (Touvron et al., 2023). Overall, these model-specific distributions show that context-aware neurons form concentrated layer-wise patterns, with their locations varying across generator models.

Mistral-7B-Instruct-v0.2 For Mistral-7B-Instruct-v0.2, the identified context-aware neurons show distinct layer-wise patterns. $\mathcal{P}_{\text{rel}}$ is concentrated around Layers 16–17 and appears again in the final layers, especially around Layers 29–31. $\mathcal{P}_{\text{irrel}}$ is also prominent around Layers 15–18, with additional occurrences around Layers 20 and 29–31. $\mathcal{P}_{\text{shared}}$ is sparse in the lower and middle layers but becomes more visible around Layers 29–31.

Qwen2.5-14B-Instruct For Qwen2.5-14B-Instruct, context-aware neurons are concentrated predominantly in the upper layers, with few neurons identified in the lower and middle layers. $\mathcal{P}_{\text{rel}}$ appears mainly after Layer 27, with visible concentrations around Layers 28–31, 36–38, and 40–47. $\mathcal{P}_{\text{irrel}}$ is also concentrated in the upper layers and forms a pronounced peak around Layers 32–33. $\mathcal{P}_{\text{shared}}$ is comparatively sparse but appears mainly around Layers 28–31 and 43–47.

LLaMA-2-13B-Chat For LLaMA-2-13B-Chat, the identified neurons are distributed across a broader range of layers. $\mathcal{P}_{\text{irrel}}$ is prominent in the early-to-middle layers, especially around Layers 8–13, and also appears in the upper layers around Layers 36–38. $\mathcal{P}_{\text{rel}}$ appears intermittently across the middle layers, including Layers 10–12, 16, and 19–22, and becomes highly concentrated in the final layers, particularly around Layers 36 and 39. $\mathcal{P}_{\text{shared}}$ is less dense overall but becomes more visible around Layers 30–33 and 36–39.

Method	NQ	ASQA	TriviaQA	SCIQ	POPQA	HotpotQA	2Wiki	AVG.
RALM	62.21	67.82	88.60	53.30	60.85	45.44	42.27	60.07
RetRobust	53.23	61.92	89.80	58.20	53.24	40.84	35.17	56.06
InstructRAG	60.62	66.98	87.48	55.30	60.25	45.33	44.61	60.08
PA-RAG	67.92	74.47	89.86	56.60	64.37	49.29	46.57	64.15
NeuRIT								
w/ Contriever	69.43	74.68	91.72	60.20	66.66	51.47	46.07	65.75
w/ Splade-v3	**69.68**	**75.31**	**92.21**	**60.80**	**67.37**	**52.25**	**46.34**	**66.28**

Table 5: Results under different retrievers with all other components fixed(SPLADE-v3 VS Contriever). Best results are in **bold** and second-best are underlined.

D.2 Robustness to Retriever Substitution.

To assess the sensitivity of the proposed neuron mining strategy to the choice of retriever, we replace SPLADE-v3 with Contriever(Izacard et al., 2021) while keeping all other components fixed. When the retriever is replaced with Contriever, the performance remains consistent with that of the SPLADE-v3 setting after re-training and evaluation. As shown in Table 5, our method outperforms existing baselines across all benchmarks, consistently achieving better performance than the strongest competitor, `PA-RAG`, on every dataset under the same setting. These results indicate that the effectiveness of the proposed approach is not tied to a specific retriever or scoring scheme, and suggest that the mined neurons capture generalizable behaviors of the generator.

Method	WOW	ELI5	FEVER
RALM	11.80	35.63	81.30
NeuRIT	**16.04**	**43.97**	**82.41**

Table 6: Performance comparison on non-QA RAG benchmarks.

D.3 Generalization Beyond QA Benchmarks.

As shown in Table 6, our method consistently outperforms `RALM` on non-QA RAG benchmarks, including WOW (Dinan et al., 2019), ELI5 (Fan et al., 2019), and FEVER (Thorne et al., 2018). These results suggest that selective neuron updates enable robust adaptation without degrading general reasoning.

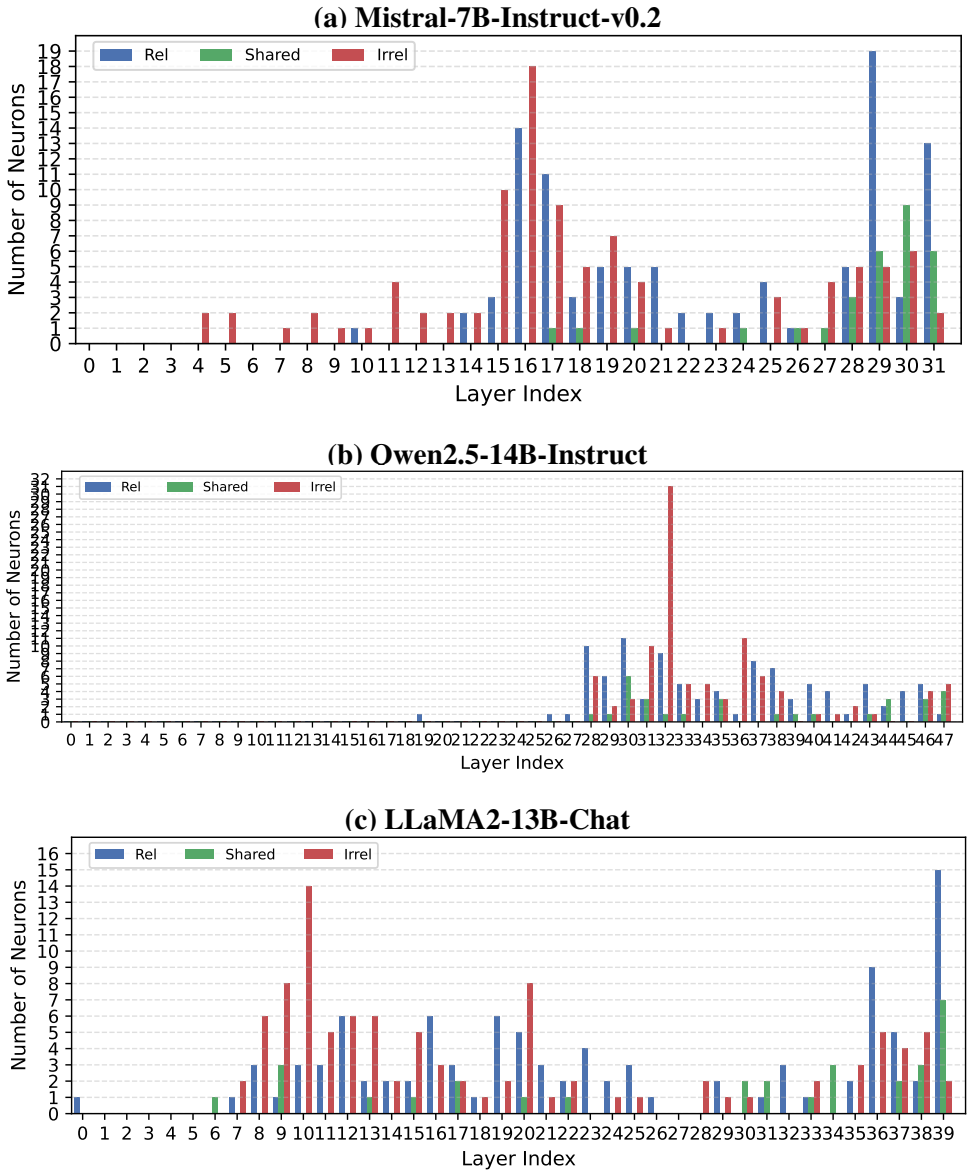

Figure 4: Layer-wise parameter and neuron distributions across models.

Examples of Relevant Attribution Set ($\tilde{\mathcal{P}}_{rel}$)	
c	Context: Arthur's Lady's Home Magazine. Arthur's Lady's Home Magazine Arthur's Home Magazine (1852–ca.1898) or Ladies' Home Magazine was an American periodical published in Philadelphia by Timothy Shay Arthur. Editors Arthur and Virginia Francis Townsend selected writing and illustrations intended to appeal to female readers. Among the contributors: Mary Tyler Peabody Mann and Kate Sutherland. In its early years the monthly comprised a selection of articles originally published in Arthur's weekly "Home Gazette." Its nonfiction stories contained occasional factual inaccuracies for the sake of a good read. → Question: Which magazine was started first Arthur's Magazine or First for Women? Proposed Answer: Arthur's Magazine
q	If the proposed answer can be derived by referring to the context, answer YES; otherwise, answer NO. The correct answer is
choices	['YES', 'NO']
gold answer	0
c	Context: The Oberoi Group. The Oberoi Group is a hotel group with its head office in Delhi. Founded in 1934, the company owns and/or operates 35 luxury hotels and two river cruise ships in six countries, primarily under its Oberoi Hotels & Resorts and Trident Hotels brands. The foundations of the Oberoi Group date back to 1934 when Rai Bahadur Mohan Singh Oberoi bought two properties in Delhi and Shimla. → Question: The Oberoi family is part of a hotel company that has a head office in what city? Proposed Answer: Delhi
q	If the proposed answer can be derived by referring to the context, answer YES; otherwise, answer NO. The correct answer is
choices	['YES', 'NO']
gold answer	0
Examples of Irrelevant Attribution Set ($\tilde{\mathcal{P}}_{irrel}$)	
c	Context: Autism. Sustained special education programs and behavior therapy early in life can help children acquire self-care, communication, and job skills, and often improve functioning and decrease symptom severity and maladaptive behaviors. Educational interventions often used include applied behavior analysis (ABA), speech and language therapy, social skills therapy, and occupational therapy. → Question: Which magazine was started first Arthur's Magazine or First for Women? Proposed Answer: Arthur's Magazine
q	If the proposed answer can be derived by referring to the context, answer YES; otherwise, answer NO. The correct answer is
choices	['YES', 'NO']
gold answer	0
c	Context: Albedo. Albedo is the measure of the diffuse reflection of solar radiation out of the total solar radiation received by an astronomical body. Enceladus has one of the highest known albedos in the Solar System, reflecting about 99% of incident radiation. → Question: The Oberoi family is part of a hotel company that has a head office in what city? Proposed Answer: Delhi
q	If the proposed answer can be derived by referring to the context, answer YES; otherwise, answer NO. The correct answer is
choices	['YES', 'NO']
gold answer	0

Table 7: Case I: Representative examples of the binary decision dataset used for Integrated Gradients-based neuron attribution. The dataset frames the task as a forced binary choice (YES/NO) to determine if the answer is strictly derivable from the context. This allows us to disentangle neurons responsible for processing relevant contexts ($\tilde{\mathcal{P}}_{rel}$) from those processing irrelevant ones ($\tilde{\mathcal{P}}_{irrel}$).

Question	Which magazine was started first Arthur's Magazine or First for Women?
Document 1	Autism. sustained special education programs and behavior therapy early in life can help children acquire self-care, communication, and job skills, and often improve functioning and decrease symptom severity and maladaptive behaviors; claims that intervention by around age three years is crucial are not substantiated. While medications have not been found to help with core symptoms, they may be used for associated symptoms, such as irritability, inattention, or repetitive behavior patterns. Section:Management.:Education. Educational interventions often used include applied behavior analysis (ABA), developmental models, structured teaching, speech and language therapy, social skills therapy, and occupational therapy. Among these approaches, interventions either treat autistic
Document 2	Albedo. back into space than what they absorb, effectively cooling the Earth. This has been a concern since arctic ice and snow has been melting at higher rates due to higher temperatures, creating regions in the arctic that are notably darker (being water or ground which is darker color) and reflects less heat back into space. This feedback loop results in a reduced albedo effect. Section:Examples of terrestrial albedo effects.:Climate and weather. Albedo affects climate by determining how much radiation a planet absorbs. The uneven heating of Earth from albedo variations between land, ice, or ocean surfaces can drive weather. Section:Examples
Document 3	Autism. social-communication skills in young children, although there is less evidence in its treatment of global symptoms. Neuropsychological reports are often poorly communicated to educators, resulting in a gap between what a report recommends and what education is provided. It is not known whether treatment programs for children lead to significant improvements after the children grow up, and the limited research on the effectiveness of adult residential programs shows mixed results. The appropriateness of including children with varying severity of autism spectrum disorders in the general education population is a subject of current debate among educators and researchers. Section:Management.:Medication. Medications may
Document 4	Albedo. the mean temperature of the planet would drop to about 0 °C. In contrast, if the entire Earth was covered by water — a so-called ocean planet — the average temperature on the planet would rise to almost 27 °C. Section:Terrestrial albedo.:White-sky, black-sky, and blue-sky albedo. For land surfaces, it has been shown that the albedo at a particular solar zenith angle "θ" can be approximated by the proportionate sum of two terms: with formula_3 being the proportion of direct radiation from a given solar angle, and formula_4 being the proportion of diffuse illumination, the actual albedo formula_5 (also called
Document 5	Autism. and the educational system are the main resources for treatment. Services should be carried out by behavior analysts, special education teachers, speech pathologists, and licensed psychologists. Studies of interventions have methodological problems that prevent definitive conclusions about efficacy. However, the development of evidence-based interventions has advanced in recent years. Although many psychosocial interventions have some positive evidence, suggesting that some form of treatment is preferable to no treatment, the methodological quality of systematic reviews of these studies has generally been poor, their clinical results are mostly tentative, and there is little evidence for the relative effectiveness of treatment options. Intensive,
Relevant summary	EOT(End of Text)

Table 8: Case II: An example of the Irrelevant Context Denoising Dataset. When provided with retrieved documents that are entirely irrelevant to the query, the model is trained to output an End-of-Text (EOT) token, enforcing the suppression of noise.

Question	Which magazine was started first Arthur's Magazine or First for Women?
Document 1	Arthur's Lady's Home Magazine. Arthur's Lady's Home Magazine Arthur's Home Magazine (1852-ca.1898) or Ladies' Home Magazine was an American periodical published in Philadelphia by Timothy Shay Arthur. Editors Arthur and Virginia Francis Townsend selected writing and illustrations intended to appeal to female readers. Among the contributors: Mary Tyler Peabody Mann and Kate Sutherland. In its early years the monthly comprised a selection of articles originally published in Arthur's weekly "Home Gazette." Its nonfiction stories contained occasional factual inaccuracies for the sake of a good read. A contemporary review judged it "gotten up in good taste and well; and is in nothing overdone. Even its
Document 2	Arthur's Magazine. Arthur's Magazine Arthur's Magazine (1844–1846) was an American literary periodical published in Philadelphia in the 19th century. Edited by Timothy Shay Arthur, it featured work by Edgar A. Poe, J.H. Ingraham, Sarah Josepha Hale, Thomas G. Spear, and others. In May 1846 it was merged into "Godey's Lady's Book". A few years later Arthur would launch a new publication entitled "Arthur's Home Magazine". Section:Further reading. Section:External links.
Document 3	Arthur (magazine). Arthur (magazine) Arthur magazine was a bi-monthly periodical that was founded in October 2002, by publisher Laris Kreslins and editor Jay Babcock. It received favorable attention from other periodicals such as "L.A. Weekly", "Print", "Punk Planet" and "Rolling Stone". "Arthur" featured photography and artwork from Spike Jonze, Art Spiegelman, Susannah Breslin, Gary Panter and Godspeed You! Black Emperor. Arthur's regular columnists included Byron Coley, Thurston Moore, Daniel Pinchbeck, Paul Cullum, Douglas Rushkoff, and T-Model Ford. "Arthur" magazine was particularly drawn to noise music, stoner metal, folk and other types of psychedelia. The first issue of "Arthur" featured an interview with
Document 4	Marianne Ehrmann. stopped the further issues. Section:Life and career.:Amaliens Erholungsstunden. "Amaliens Erholungsstunden" was the first women's magazine, and it was self-published monthly by Marianne Ehrmann under her own name. She sent advance notices to acquaintances, friends and potential patrons to distribute these in the circle, and the first six editions resulted in an increase of subscribers. The debts contracted by the self-publishing were indeed paid off, but in all a financial disaster, that's why in January 1791 Marianne and her husband started a cooperation with the publishing house "J. G. Cottaische Verlagsbuchhandlung" in Tübingen. Marianne Ehrmann thought to concentrate on the journalistic
Document 5	Arthur Jewitt. "The Northern Star, or Yorkshire Magazine," a monthly register of arts, biography, statistics, manufactures, &c., which ran to three volumes, 1817–18. On 1 January 1818 he brought out the first number of "The Sylph, or Lady's Magazine for Yorkshire, Derbyshire, and the adjoining Counties." "The Lincoln and Lincolnshire Cabinet and Annual Intelligencer" appeared at Lincoln during 1827–9. His "Matlock Companion," 1835, and "Derbyshire Gems" were very popular. His "Handbook of Practical Perspective," 1840, and his "Handbook of Geometry," 1842, were adopted by the committee of council on education. He contributed mathematical papers to the "British Diary" and to the "Lady's"
Relevant summary	The document mentions "Arthur's Magazine" as an American literary periodical published from 1844 to 1846. It was edited by Timothy Shay Arthur and later merged into "Godey's Lady's Book" in 1846. It also references "Arthur's Home Magazine" starting in 1852, but no start date or details are given for "First for Women." There is no mention of "First for Women" magazine's start date or any comparison to Arthur's Magazine.

Table 9: Case III: An example of the Relevant Summary Dataset ($\mathcal{D}_{RS}$). The external LLM synthesizes a concise summary using only the query-aligned documents, while effectively filtering out unrelated information.