SCoNE : Selective Context-aware Neuron Editing for Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) is highly sensitive to retrieval noise: when retrieved documents mix informative and irrelevant context, LLMs are easily distracted, leading to hallucinations. To overcome this, we propose SCoNE (Selective Context-aware Neuron Editing), a training-free model editing approach that improves retrieval noise robustness by selectively strengthening context-aware FFN neurons that are identified by both high attribution and high cross-input variability. SCoNE requires only a small number of mining samples, no fine-tuning, and no inference-time overhead. Across various knowledge-intensive question-answering benchmarks and two LLM backbones, SCoNE consistently outperforms competitive baseline methods.

1 Introduction

While Large Language Models (LLMs) have achieved remarkable success, they remain prone to hallucinations in knowledge-intensive tasks (Wang and Yu, 2025; Huang et al., 2025). Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) mitigates this by grounding outputs in externally retrieved evidence. However, the effectiveness of RAG is heavily dependent on the quality of retrieved documents, which retrieval systems cannot always guarantee. In realistic settings, retrievers return a mixture of relevant, partially relevant, and irrelevant documents for the same query, and LLMs are known to be easily distracted by such *retrieval noise*, often degrading rather than improving their answers (Yoran et al., 2024; Shi et al., 2023).

Existing approaches to retrieval noise robustness span several paradigms, including prompt engineering (Zhou et al., 2023), retrieved-context refinement such as reranking and compression (Glass et al., 2022; Xu et al., 2023), and fine-tuning. While solutions such as introducing additional modules (*e.g.,* reranker, compressor) are flexible, they introduce additional components into the RAG pipeline, which can lead to cascading errors across each stage (Asai et al., 2024; Yoran et al., 2024) and substantial inference-time latency (An et al., 2025). In contrast, directly fine-tuning the generator to be robust against retrieval noise avoids such pipeline overhead and has therefore emerged as a promising direction. Yoran et al. (2024) train the generator on mixtures of relevant and irrelevant contexts, and Wu et al. (2025) align it via multi-perspective preference optimization. However, fine-tuning-based methods inherit the well-known drawbacks of gradient-based adaptation: catastrophic forgetting, substantial compute requirements, and the need for carefully curated training data. A natural question arises: *can retrieval noise robustness be achieved without retraining the model?*

Model editing offers a promising alternative paradigm. By directly modifying a small number of parameters, editing methods provide fine-grained control over model behavior without the cost of fine-tuning. However, existing model editing methods fundamentally assume that the target knowledge to be edited is known in advance (Meng et al., 2022a,b). This assumption does not hold in Retrieval-Augmented Generation (RAG), where retrieved contexts are inherently open-ended and dynamically vary across queries. In RAG, the model cannot anticipate which facts will appear in the retrieved documents at inference time. Hence, rather than fixing a specific parametric knowledge, RAG requires an adaptive approach to constrain the model's behavior in response to whatever context is retrieved at inference, regardless of its specific content. This challenge is further complicated by the realistic nature of retrieval itself; even for a single query, some documents directly support the answer, others are partially relevant, and others are entirely irrelevant. For this, a method that not only makes the model responsive to context, but selectively responsive to a context (*i.e.,* engaging with

informative documents while remaining unaffected by noisy ones) is necessarily required.

Building on this insight, we propose **SCoNE** (**S**elective **Co**ntext-aware **N**euron **E**diting) for RAG, which is a model editing method to enhance robustness for retrieval noise. SCoNE identifies *selectively context-aware neurons* by jointly requiring high attribution and high cross-input variability, and strengthens them at inference time. Our method requires only 100 mining samples from a single dataset (HotpotQA), no fine-tuning, and no inference-time overhead beyond standard RAG. Across various benchmarks, SCoNE consistently outperforms strong competitive baselines, demonstrating that lightweight editing, when guided by the right neuron selection criterion, can match or exceed the effectiveness of heavyweight fine-tuning pipelines. Similarly, Shi et al. (2024a) adopt a knowledge-agnostic approach but identify context-aware neurons based on attribution strength alone. While this criterion is effective in their single-context setting, RAG presents multiple retrieved documents containing both informative and distracting evidence simultaneously. Here, a neuron may receive high attribution simply because it responds broadly to any retrieved content, potentially reflecting sensitivity to surface-level patterns rather than informative evidence. Thus, attribution strength alone is insufficient for identifying neurons that specifically mediate useful evidence utilization in noisy retrieval settings. Our redefinition of context-aware neurons addresses this gap by complementing attribution strength with cross-input variability, making the criterion more suitable for RAG's heterogeneous, multi-document setting.

2 Method

Problem Setup. Consider a neuron mining dataset of size N, sampled from the training split of HotpotQA (Yang et al., 2018), $\mathcal{D} = \{(q, \mathcal{C}, a)_t\}_{t=1}^{N}$, where the t-th instance consists of a query q_t, its ground-truth answer a_t, and an associated context set $\mathcal{C}_t = \{c_m^g\}_{m=1}^{M} \cup \{c_k^d\}_{k=1}^{K}$. Here, $\{c_m^g\}_{m=1}^{M}$ denote the gold contexts that directly support a_t, while $\{c_k^d\}_{k=1}^{K}$ are distractor contexts that do not entail the answer. Our goal is to characterize, for each instance, how individual FFN neurons engage with such mixed evidence during inference. To this end, we define two complementary measures, *attribution* and *variability*, computed over the course of inference.

2.1 Neuron Mining

Attribution. Prior work has shown that factual knowledge is localized in specific FFN neurons (Dai et al., 2022; Geva et al., 2021). We hypothesize that neurons responsible for processing retrieved contextual information also reside in FFNs, and seek to identify those that engage with retrieved evidence during RAG inference. Following Shi et al. (2024a), we estimate the attribution of each FFN neuron n via Integrated Gradients (Sundararajan et al., 2017), but extend their single-context formulation to the multi-context RAG setting where gold and distractor evidence co-occur. For each instance $(q, \mathcal{C}, a)_t \in \mathcal{D}$, we quantify how individual FFN neurons contribute to the model's prediction when the query is presented together with its full context set $\mathcal{C}_t = \{c_m^g\}_{m=1}^{M} \cup \{c_k^d\}_{k=1}^{K}$, which contains both gold and distractor contexts. Specifically, we formulate the attribution score calculation as follows: let $v(q_t)$ denote the activation of neuron n when the model is given the query alone, and let $v(q_t, \mathcal{C}_t)$ denote its activation when the query is augmented with the full context set $\mathcal{C}_t$. The attribution score is then defined as follows:

$$\mathrm{Attr}(n;\, q_t, \mathcal{C}_t) = \big(v(q_t, \mathcal{C}_t) - v(q_t)\big) \times \int_{\alpha=0}^{1} \frac{\partial\, P(a \mid q_t, \mathcal{C}_t, v_\alpha)}{\partial\, v_\alpha}\, d\alpha, \tag{1}$$

where $v_\alpha = v(q_t) + \alpha\big(v(q_t, \mathcal{C}_t) - v(q_t)\big)$ linearly interpolates between the query-only and the query-plus-context activations for $\alpha \in [0, 1]$. In practice, the integral is approximated by a 20-step Riemann sum.

Variability. Attribution alone, however, cannot distinguish between two qualitatively different neuron behaviors: neurons that selectively respond to specific contexts and neurons that activate uniformly regardless of context content. Only the former captures the content-level selectivity required in the heterogeneous multi-document setting of RAG. To capture this context selectivity, we measure how a neuron's attribution varies across different query-context instances. Concretely, let $\mathrm{Attr}^{(t)}(n_j^l)$ denote the attribution of the j-th intermediate neuron in the l-th FFN layer for the t-th instance $(q, \mathcal{C}, a)_t \in \mathcal{D}$. The variability score is defined as the deviation of the current attribution from its running average over the preceding W

	NQ	ASQA	SCIQ	TriviaQA	HQA	TruthfulQA	PopQA	Avg.
Llama-3-8B-Instruct								
RAG (Lewis et al., 2020)	62.81	68.78	54.10	88.65	46.55	4.90	60.17	55.14
RetRobust (Yoran et al., 2024)	53.75	60.86	**59.10**	89.88	40.98	2.08	53.93	51.51
PA-RAG (Wu et al., 2025)	**68.06**	73.73	56.80	90.18	50.41	3.79	64.19	58.17
CAD (Shi et al., 2024b)	64.12	68.99	48.50	87.81	46.77	3.55	62.21	54.56
IRCAN (Shi et al., 2024a)	64.65	71.41	54.00	89.67	50.11	5.51	63.65	57.00
SCoNE (Ours)	66.44	**73.84**	57.10	**90.66**	**52.27**	**6.36**	**65.57**	**58.89**
Qwen-2.5-7B-Instruct								
RAG (Lewis et al., 2020)	62.25	69.09	54.90	86.98	45.54	**6.24**	58.54	54.79
RetRobust (Yoran et al., 2024)	60.80	67.51	54.10	86.74	44.46	5.75	57.52	53.84
PA-RAG (Wu et al., 2025)	60.45	68.35	52.90	87.54	**48.00**	6.00	56.65	54.27
CAD (Shi et al., 2024b)	61.83	69.62	51.50	85.62	43.88	4.04	58.84	53.62
IRCAN (Shi et al., 2024a)	62.25	68.78	54.20	87.28	45.86	6.00	58.82	54.74
SCoNE (Ours)	**63.24**	**69.94**	**56.20**	**88.16**	47.00	5.63	**60.04**	**55.74**

Table 1: Accuracy comparison across QA benchmarks using Llama-3-8B-Instruct (top) and Qwen-2.5-7B-Instruct (bottom). **Bold** scores are best in each dataset, and underlined scores are the second-best results.

instances:

$$V^{(t)}(n_j^l) = \left| \mathrm{Attr}^{(t)}(n_j^l) - \frac{1}{W} \sum_{m=t-W}^{t-1} \mathrm{Attr}^{(m)}(n_j^l) \right|, \tag{2}$$

where W specifies the number of preceding instances included in the running average. We compute this score over a fixed random traversal order of $\mathcal{D}$, so that the sliding window reflects the neuron's behavior on an unbiased sample of recently observed instances rather than any task- or label-specific ordering.

High-Attribution and Variability Neuron Selection. Neurons with both high attribution and high variability are considered context-aware: they contribute strongly to the current context while responding differently across inputs. Hence, we select the neurons as follows: For each sample $(q, \mathcal{C}, a)_t$, we construct $\mathcal{A}^{(t)}$ and $\mathcal{B}^{(t)}$, containing the top-50 neurons ranked by attribution and variability, respectively, restricted to neurons with positive attribution scores.[1] Their intersection $\mathcal{C}^{(t)} = \mathcal{A}^{(t)} \cap \mathcal{B}^{(t)}$ forms the locally selected neurons for sample t. We then aggregate the selection frequency of each neuron across all samples and choose the top-k most frequent neurons as the final context-aware neuron set.

2.2 Neuron Enhancement

Once context-aware neurons are identified, we amplify their contribution at inference time to better leverage informative retrieved evidence. For each selected neuron n_i^l, we scale its corresponding FFN weight as:

$$\hat{W}(n_i^l) = \alpha \cdot W(n_i^l)$$

where α controls the enhancement strength.

3 Experimental Setup

Neuron Mining Dataset. We randomly sample 100 instances from the HotpotQA (Yang et al., 2018) training split for neuron mining. We set the number of gold-content $M = 2$, and the distractor context $K = 8$.

Evaluation Dataset. We use the dev splits provided by BERGEN (Rau et al., 2024) for NQ (Kwiatkowski et al., 2019), ASQA (Stelmakh et al., 2022), SCIQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018), TruthfulQA (Lin et al., 2022), PopQA (Mallen et al., 2023). All retrieval documents are sourced from the KILT (Petroni et al., 2021) Wikipedia dump[2], and we retrieve top-5 documents per question using SPLADE-v3 (Lassance et al., 2024).

Baselines. We compare SCoNE against representative RAG baselines: RAG (Lewis et al., 2020); RetRobust (Yoran et al., 2024) and PA-RAG (Wu et al., 2025) for generator fine-tuning; and CAD (Shi et al., 2024b) and IRCAN (Shi et al., 2024a) for inference-time intervention at the decoding and parameter level, respectively.

[1] Restricting to positive attribution prevents neurons whose variability stems from transitions between negative and near-zero attribution, whose overall contribution remains negligible, from being selected.

[2] https://huggingface.co/datasets/facebook/kilt_wikipedia

Llama-3-8B-Instruct	Relevant			Irrelevant		
	NQ	SCIQ	HQA	NQ	SCIQ	HQA
RAG (Lewis et al., 2020)	78.04	73.61	72.75	4.43	**8.36**	15.16
PA-RAG (Wu et al., 2025)	**85.29**	**78.60**	**81.26**	2.04	5.69	13.43
IRCAN (Shi et al., 2024a)	80.44	73.75	76.88	4.09	7.69	18.02
SCoNE (Ours)	82.00	78.03	79.53	**6.81**	8.03	**19.59**

Table 2: The comparison of accuracy on Relevant and Irrelevant subsets.

Implementation Details. Our experiments are performed using the RAG framework provided by BERGEN (Rau et al., 2024), which offers a realistic RAG pipeline. We use Llama-3-8B-Instruct (Touvron et al., 2023) and Qwen-2.5-7B-Instruct (Yang et al., 2024) as the generator LLM, and SPLADE-v3 (Lassance et al., 2024) as the retriever. For retrieval, we use the KILT Wikipedia dump[3], preprocessed into non-overlapping 100-word chunks, and retrieve five documents per question. All experiments are conducted on a single NVIDIA H200 GPU. For fair comparison, both IRCAN and our method identify neurons from the same neuron mining dataset $\mathcal{D}$. Details of $\mathcal{D}$ are provided in A.1. We set the enhancement strength $\alpha = 7$, select the top-k context-aware neurons where $k = 5$, and the window size $W = 3$.

4 Results

Main Results. Table 1 reports the main results. SCoNE achieves the best overall performance with Llama-3-8B-Instruct, ranking first on five of seven datasets and improving over RAG by 3.75% on average. Compared to fine-tuning baselines, SCoNE outperforms RetRobust by 7.38% and remains within 0.72% of PA-RAG despite requiring no additional training. Against intervention-based baselines, SCoNE surpasses CAD on all datasets and improves over IRCAN by up to 3.1%, suggesting that our variability-based criterion identifies neurons more selectively responsive to retrieved context than attribution alone. A similar trend holds with Qwen-2.5-7B-Instruct as the generator, where SCoNE consistently outperforms IRCAN across most benchmarks. SCoNE achieves the best average accuracy, demonstrating its effectiveness across different generators.

Relevant vs. Irrelevant Context Analysis. To examine whether our method exhibits context-dependent behavior with retrieved contexts, we divide each evaluation set into two subsets: *Relevant*, where at least one retrieved document contains the gold answer, and *Irrelevant*, otherwise.

[3]https://huggingface.co/datasets/kilt_wikipedia

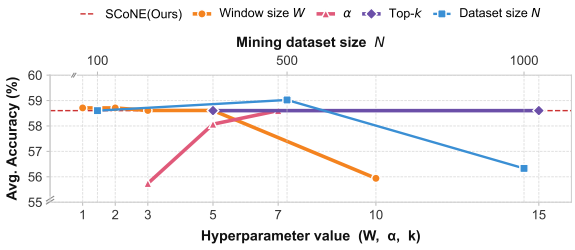

Figure 1: Ablation study on key hyperparameters using Llama-3-8B-Instruct. Results are averaged across NQ, SCIQ, and HotpotQA.

As shown in Table 2, SCoNE consistently improves over RAG and IRCAN on the relevant subset across all datasets. While PA-RAG achieves the highest accuracy on the relevant subset, SCoNE demonstrates stronger robustness under irrelevant contexts, outperforming all baselines, including vanilla RAG, on NQ and HQA. Notably, SCoNE surpasses IRCAN on both the Relevant and Irrelevant subsets, suggesting that incorporating cross-input variability beyond attribution strength helps identify neurons that selectively engage with informative evidence.

Hyperparameter Analysis. We conduct ablation studies on the enhancement strength α, context window size W, neuron mining dataset size N, and the number of selected neurons k using Llama-3-8B-Instruct. The result is shown in Figure 1.[4] Performance improves monotonically with α, peaking at $\alpha = 7$, and remains stable across small-to-moderate W, N, and k. Overall, while extreme hyperparameter values (*e.g.*, $W = 10$ or $N = 1000$) degrade performance, the default SCoNE configuration denoted as *SCoNE (Ours)* consistently achieves the best or near-best performance, validating our design choices.[5]

5 Conclusion

We present SCoNE, a framework that improves RAG robustness to retrieval noise by selectively enhancing context-aware neurons, identified through attribution strength and cross-input variability. Experiments across various benchmarks show that SCoNE matches or surpasses strong baselines, demonstrating that variability-based neuron mining provides a practical criterion.

[4]More detailed results for each hyperparameter setting are provided in A.4, A.5, A.6.

[5]SCoNE (Ours) corresponds to the default configuration with $\alpha = 7$, $W = 3$, $N = 100$, and $k = 5$.

6 Limitations

While the selected neurons transfer effectively across diverse benchmarks, several limitations remain. In this work, neuron mining is performed using samples from HotpotQA only, and it remains unclear how the characteristics of the mining dataset influence the selected neuron set and downstream behavior. For example, using more challenging QA datasets may lead to different neuron distributions and transfer properties. In addition, our current setting relies on contexts containing both gold-content and distractor-content. Therefore, it remains unclear how neuron selection would differ under cleaner retrieval settings containing only gold supporting documents. Investigating how mining dataset composition and retrieval conditions affect neuron mining and robustness remains an important direction for future work.

References

Yuwei An, Yihua Cheng, Seo Jin Park, and Junchen Jiang. 2025. Hyperrag: Enhancing quality-efficiency tradeoffs in retrieval-augmented generation with reranker kv-cache reuse. *arXiv preprint arXiv:2504.02921*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

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, Dublin, Ireland. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

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.

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.

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. *arXiv preprint arXiv:2403.06789*.

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.

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, Dublin, Ireland. Association for Computational Linguistics.

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, Toronto, Canada. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2523–2544, Online. Association for Computational Linguistics.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. 2024. 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. 2024a. Ircan: Mitigating knowledge conflicts in llm generation via identifying and reweighting context-aware neurons. *Advances in Neural Information Processing Systems*, 37:4997–5024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024b. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

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. 2025. 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.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv e-prints*, pages arXiv–2412.

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)*.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Neuron Mining Dataset Setting

For neuron mining, we use the first 100 samples from the HotpotQA training distractor split (Yang et al., 2018). To simulate a realistic RAG setting where multiple retrieved documents are provided as input, we convert each sample into the context-provided prompt format shown in Table 3, where all documents in the sample are directly used as the input context without an additional retriever. The corresponding prompt format without retrieved context is shown in Table 4.

Input Format With Retrieved Context

System: Your task is to extract relevant information from provided documents and to answer questions as briefly as possible.
User:
Background:
Document 1: [title] [sentences]
Document 2: [title] [sentences]
⋮
Document N: [title] [sentences]
Question: [question]

Table 3: Input prompt format with retrieved context, used for attribution score computation.

Input Format Without Retrieved Context

System: Answer the questions as briefly as possible.
User: Question: [question]

Table 4: Input prompt format without retrieved context, used for attribution score computation.

A.2 Inference-Time RAG Prompt

The prompt format used for inference-time RAG generation is shown in Table 5. The "Background" section consists of the top-5 documents retrieved by the retriever.

A.3 Evaluation Metric

We evaluate our methods using one metric: Accuracy. Accuracy follows the Match score, which measures whether the gold label appears as a substring in the generated output.

Inference Input Format

System: You are a helpful assistant. Your task is to extract relevant information from provided documents and to answer questions as briefly as possible.
User:
Background:
Document 1: ...
Document 2: ...
⋮
Document 5: ...
Question: [question]

Table 5: Input prompt format used for inference-time RAG.

	NQ		SCIQ		HQA	
α	k=5	k=15	k=5	k=15	k=5	k=15
3	63.80	63.27	53.50	54.00	49.91	49.57
5	65.81	65.67	56.20	58.00	52.20	52.66
7	66.44	66.02	57.10	57.70	52.27	52.09

Table 6: Effect of enhancement strength α and top-k neuron mining.

A.4 Effect of Enhancement Strength and Number of Selected Neurons

We additionally evaluate different enhancement strengths $\alpha \in \{3, 5, 7\}$ and top-k values $k \in \{5, 15\}$ on KILT-NQ, KILT-HotpotQA, and SCIQ using Llama-3-8B-Instruct. Larger enhancement strengths generally lead to better performance, with $\alpha = 7$ achieving the strongest overall results. We also observe that $k = 5$ and $k = 15$ yield comparable performance, suggesting that the neurons most important for selective context utilization are already concentrated within a small set of top-ranked neurons.

A.5 Effect of Context Window Size

We vary the context window size $W \in \{1, 2, 3, 5, 10\}$ used to compute attribution variability. As shown in Figure 2, performance remains stable across small and moderate window sizes ($W = 1, 2, 3, 5$), while a large window ($W = 10$) consistently degrades performance across datasets.

We additionally observe that the mined neuron sets are highly similar across different window sizes. In particular, the selected neurons for $W = 1$ and $W = 2$ are identical, and those for $W = 3$ and $W = 5$ are also identical. Even for $W = 10$, the selected neuron set differs from the other settings

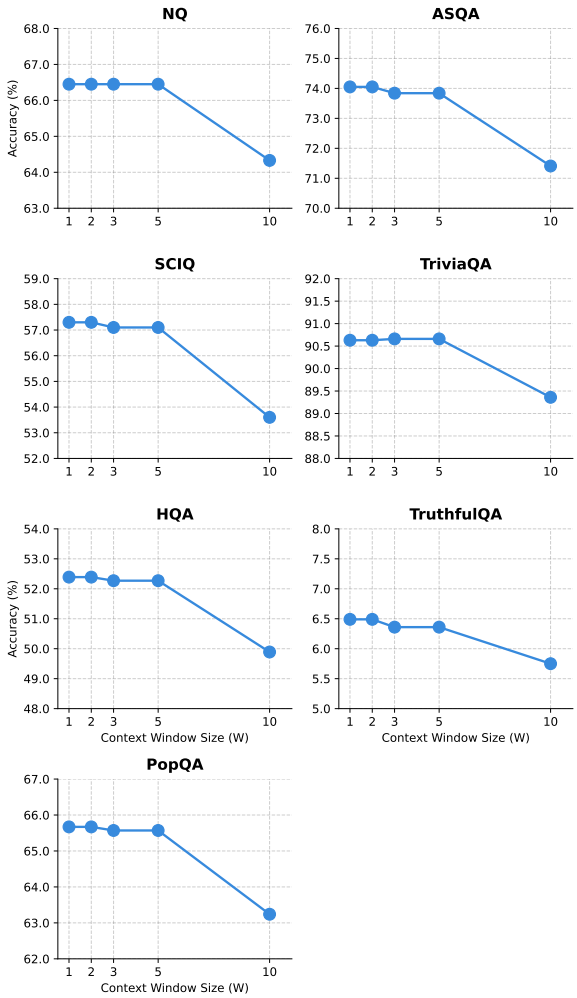

Figure 2: Performance sensitivity to context window size W.

by at most one neuron. Despite these highly similar neuron sets, performance differences remain relatively small for practical window sizes, with degradation mainly observed at $W = 10$.

A.6 Effect of Neuron Mining Dataset Size

We additionally analyze the effect of neuron-mining dataset size by selecting neurons using $N \in \{100, 500, 1000\}$ samples from HotpotQA on Llama-3-8B-Instruct. Table 7 reports the performance on KILT-NQ, SCIQ, KILT-HotpotQA.

We observe that increasing the number of neuron-mining samples does not necessarily improve downstream performance. In particular, $N = 100$ and $N = 500$ yield comparable results, while performance drops when using $N = 1000$, despite requiring substantially more mining data. The selected neurons also exhibit considerable overlap across different sample sizes. These results suggest that context-aware neurons can be identified with relatively small mining sets.

# Samples	NQ	SCIQ	HQA	Overlap
100	66.44	57.10	52.27	–
500	66.94	57.50	52.63	3/5
1000	64.61	54.10	50.28	3/5

Table 7: Effect of neuron mining dataset size on HotpotQA using LLaMA-3-8B-Instruct. Overlap indicates the percentage of shared neurons compared to the 100-sample setting.