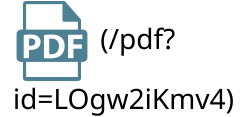


← Go to **ACL ARR 2025 October** homepage (/group?id=aclweb.org/ACL/ARR/2025/October)

5 Versions ▾

# CachePrune: Neural-Based Attribution Defense Against Indirect Prompt Injection Attacks



*Rui Wang* (/profile?id=~Rui\_Wang25), *Junda Wu* (/profile?id=~Junda\_Wu1),  
*Yu Xia* (/profile?id=~Yu\_Xia9), *Tong Yu* (/profile?id=~Tong\_Yu3),  
*Ruiyi Zhang* (/profile?id=~Ruiyi\_Zhang3),  
*Ryan A. Rossi* (/profile?id=~Ryan\_A.\_Rossi2),  
*Subrata Mitra* (/profile?id=~Subrata\_Mitra1), *Lina Yao* (/profile?id=~Lina\_Yao2),  
*Julian McAuley* (/profile?id=~Julian\_McAuley1)

07 Oct 2025 (modified: 17 Mar 2026) ACL ARR 2025 October Submission

October, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers

Revisions (/revisions?id=LOgw2iKmv4)

CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

## Abstract:

Large Language Models (LLMs) are susceptible to indirect prompt injection attack, in which the model inadvertently responds to task messages injected within the prompt context. This vulnerability stems from LLMs' inability to distinguish between data and instructions within a prompt. In this paper, we propose CachePrune that defends against this attack by identifying and pruning task-triggering neurons from the KV cache of the input prompt context. By pruning such neurons, we encourage the LLM to interpret the input prompt context purely as data rather than any cues for instruction following. To identify these neurons, we introduce a neural attribution mechanism guided by a preferential attribution loss, which enables effective attribution with only a few samples while preserving response quality after pruning. Notably, our approach does not impose additional formatting on the prompt or introduce extra test-time LLM calls. Experiments show that CachePrune can significantly reduce attack success rates while not compromising the response quality.

**Paper Type:** Long

**Research Area:** Efficient/Low-Resource Methods for NLP

**Research Area Keywords:** LLM, Neural Attribution

**Contribution Types:** NLP engineering experiment, Approaches to low-resource settings

**Languages Studied:** English

**Previous URL:** /forum?id=jsJrYWNTv3 (/forum?id=jsJrYWNTv3)

**Explanation Of Revisions PDF:** pdf (/attachment?id=LOgw2iKmv4&name=explanation\_of\_revisions\_PDF)

**Reassignment Request Area Chair:** Yes, I want a different area chair for our submission

**Reassignment Request Reviewers:** Yes, I want a different set of reviewers

**Justification For Not Keeping Action Editor Or Reviewers:** Feedbacks in the previous round asked more about the setting. However, my setting is following the previous works. Therefore, I think they may not be familiar with this area.

**Preprint:** yes

**Preprint Status:** There is a non-anonymous preprint (URL specified in the next question).

**Existing Preprints:** <https://arxiv.org/abs/2504.21228> (<https://arxiv.org/abs/2504.21228>)

**Preferred Venue:** ACL

**Consent To Share Data:** yes

**Consent To Share Submission Details:** On behalf of all authors, we agree to the terms above to share our submission details.

**A1 Limitations Section:** This paper has a limitations section.

**A2 Potential Risks:** Yes

**A2 Elaboration:** 5

**B Use Or Create Scientific Artifacts:** Yes

**B1 Cite Creators Of Artifacts:** Yes

**B1 Elaboration:** 4

**B2 Discuss The License For Artifacts:** No

**B2 Elaboration:** All public

**B3 Artifact Use Consistent With Intended Use:** No

**B3 Elaboration:** All public

**B4 Data Contains Personally Identifying Info Or Offensive Content:** No

**B4 Elaboration:** All public

**B5 Documentation Of Artifacts:** No

**B5 Elaboration:** All public

**B6 Statistics For Data:** Yes

**B6 Elaboration:** 4

**C Computational Experiments:** Yes

**C1 Model Size And Budget:** Yes

**C1 Elaboration:** I report the models used and their sizes in section 4.

**C2 Experimental Setup And Hyperparameters:** Yes

**C2 Elaboration:** 4

**C3 Descriptive Statistics:** Yes

**C3 Elaboration:** 4

**C4 Parameters For Packages:** Yes

**C4 Elaboration:** Appendix

**D Human Subjects Including Annotators:** No

**D1 Instructions Given To Participants:** N/A

**D2 Recruitment And Payment:** N/A

**D3 Data Consent:** N/A

**D4 Ethics Review Board Approval:** N/A

**D5 Characteristics Of Annotators:** N/A

**E Ai Assistants In Research Or Writing:** Yes

**E1 Information About Use Of Ai Assistants:** No

**E1 Elaboration:** Only use ChatGPT to polish some writing.

**Author Submission Checklist:** yes

**Association For Computational Linguistics - Blind Submission License Agreement:**  On behalf of all authors, I agree

**Submission Number:** 3650

Filter by reply type... ▼

Filter by author... ▼

Search keywords...

Sort: Newest First

-

=

≡

Everyone

Submission3650...

Submission3650 Area...

Submission3650 Authors

13 / 13 replies shown

Submission3650...

Program Chairs

Submission3650...

Submission3650...

Submission3650...

Submission3650...

Submission3650...

✕

### Meta Review of Submission3650 by Area Chair vGH7

Meta Review by Area Chair vGH7 04 Dec 2025, 00:56 (modified: 17 Mar 2026, 09:30)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=05FGE2ZYzu)

**Metareview:**

This paper proposes CachePrune, a defense method that identifies and prunes task-triggering neurons from the KV cache of the input prompt context. Reviewers agree that the approach is novel, lightweight, and practically useful, with clear motivation and strong empirical results. The attribution loss is well-justified, and the mechanistic analysis provides additional insight.

The main concerns involved comparisons to training-based defenses, potential impact on reasoning ability, and the stability of attribution with small sample sets. The authors addressed these issues with additional experiments and clarifications.

Overall, the paper presents a promising and original idea with generally convincing empirical results. I recommend Borderline Conference.

**Summary Of Reasons To Publish:**

This paper introduces a novel and lightweight KV-cache-level defense with clear motivation, solid empirical evidence, and meaningful mechanistic insights, making it a promising contribution to prompt-injection robustness for ACL audiences.

**Summary Of Suggested Revisions:**

The paper would benefit from incorporating the clarifications, analyses, and extended discussions provided during the rebuttal into a future revised version, which would help further strengthen the presentation and make the contributions clearer to readers.


**Overall Assessment:** 3.5 = Borderline Conference

**Reported Issues:**  No

**Publication Ethics Policy Compliance:** I did not use any generative AI tools for this review

**Kind Request for Reviewer Feedback on Rebuttal**

Author-Editor Confidential Comment

by Authors ( Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))


 25 Nov 2025, 12:15  Program Chairs, Senior Area Chairs, Area Chairs, Authors

**Comment:**


We believe that our rebuttal has sufficiently addressed the concerns raised in the reviews, particularly those by reviewers 1Lz7 and mfWA. If possible, we would greatly appreciate it if you could encourage them to engage with our responses. Their feedback would be valuable in ensuring a fair and informed evaluation. Thank you so much for your consideration.

**Invitation to Review Author Rebuttal and Updates**

Official Comment

by Authors ( Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

 24 Nov 2025, 12:10 (modified: 17 Mar 2026, 09:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

 Revisions (/revisions?id=wqS98p54Vf)

**Comment:**

We sincerely thank all reviewers for their time and thoughtful feedback. We have carefully addressed the concerns raised in the initial reviews and provided detailed responses in the rebuttal.

We kindly invite reviewers to take a look at our responses and updates, and we welcome any further comments or suggestions.



## Follow-Up

Author-Editor Confidential Comment

by Authors (👁️ Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

📅 02 Dec 2025, 00:13 (modified: 07 Dec 2025, 21:06)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Authors 📄 Revisions (/revisions?id=jXiaYaxUwq)

### Comment:

We would like to gently follow up on our submission. We wanted to note that we have provided detailed responses to all raised concerns. This includes presenting results comparing CachePrune to DPO-based fine-tuning (Reviewer 1Lz7, mfWA), though we have explained (e.g., Section 1, 3) that our CachePrune is lightweighted and requires much less computation than fine-tuning. We also add experiments confirming that CachePrune preserves response quality under CoT-enabled settings (Reviewer qbZK).

Thank you again for your time and effort in managing the review process.

## Official Review of Submission3650 by Reviewer 1Lz7

Official Review by Reviewer 1Lz7 📅 12 Nov 2025, 20:49 (modified: 17 Mar 2026, 09:30)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer 1Lz7, Authors, Commitment Readers

📄 Revisions (/revisions?id=fyqboPGZcN)

### Paper Summary:

The paper proposes a test-time defense against indirect prompt injection attacks on LLMs. The method identifies neurons in the KV cache that disproportionately contribute to poisoned behavior, and prunes them while preserving neurons important for clean responses. The authors introduce a preferential attribution loss, demonstrating its relationship to DPO-style preference optimization. Experiments show the proposed approach reduces attack-success rates without degrading output quality across various datasets and models.

### Summary Of Strengths:

1. The preferential attribution loss and neuron-level pruning strategy are well-connected to prior theoretical frameworks such as DPO
2. The approach of using most-probable clean and poisoned responses (greedy) reduces the number of required samples and avoids extra LLM calls, which is a practical advantage over previous defenses approaches
3. The authors demonstrations that only a few initial tokens can trigger poisoned behavior, which provide strong intuition for why the method works

### Summary Of Weaknesses:

1. There is limited comparison to training-based defenses: The authors acknowledge that they exclude many fine-tuning approaches despite these being leading methods, reducing the completeness of the empirical comparison.
2. Greedy decoding is used as a proxy for the true most-probable outputs. However, the paper notes this is an approximation but does not fully quantify its impact on attribution quality.
3. While some model variety is included (e.g., Mistral, LLaMA3), the method's reliance on neuron-level attribution may not transfer uniformly across significantly different architectures.
4. Although the adaptive attacks are mentioned in the appendix, the main text does not deeply analyze how an adversary might circumvent cache-level pruning.

### Comments Suggestions And Typos:

1. Adding training-based defenses (even if computationally expensive) would provide a fuller picture of trade-offs and practical deployment considerations.
2. More detailed exploration of how pruning affects model internal representations, latency, and token probabilities would strengthen the argument for safety and reliability.
3. Include models with different internal mechanisms (e.g., MoE) can help test generalizability of the approach

Typos: L572: exhibites → exhibits L348: aggregration -> aggregation Figure 3: Posioned -> Poisoned

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Soundness:** 2.5

**Excitement:** 2.5

**Overall Assessment:** 3 = Findings: I think this paper could be accepted to the Findings of the ACL.

**Limitations And Societal Impact:**

They've partially addressed limitations but have not really engaged with societal impacts.

**Ethical Concerns:**

There are no concerns with this submission

**Needs Ethics Review:** No

**Reproducibility:** 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

**Publication Ethics Policy Compliance:** I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



## Reply to Reviewer 1Lz7

Official Comment

by Authors (👁️ Ryan A. Rossi (/profile?id=~Ryan\_A\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

📅 23 Nov 2025, 04:36 (modified: 17 Mar 2026, 09:30)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 1Lz7, Commitment Readers

📄 Revisions (/revisions?id=IoTuzdDvF)

**Comment:**

Thank you very much for your valuable and insightful feedback!

**1. Comparing to finetuning-based approach**

Please refer to the other reply for additional experiments on finetuning with DPO.

**2. Approximating with the true most-probable outputs with greedy decoding**

Since it is computational infeasible to sample the true most-probable response, prior work commonly adopts greedy decoding as a **standard approximation, e.g., some well cited and high impact works [1–4]**. Additionally, it is the **default decoding strategy in the widely used Huggingface Transformers [5]**. Note that, despite approximating with greedy decoding, our CachePrune can still significantly reduce the Attack Success Rate while maintaining the response quality (Table 1).

**3. Reliance on neuron-level attribution affects generalization different architectures (e.g., MOE).**

Our method is **applicable to any autoregressive LLM (MOE or not)** that (i) encodes the prompt into neural activations and (ii) generates outputs conditioned on these activations. Specifically, CachePrune attributes the model outputs to neural activations via our proposed attribution loss (eq 13), prunes neurons that are responsible for poisoned responses while not degrading the quality of clean

responses. In Table 1, our approach **consistently** and **substantially** reduces the Attack Success Rate across LLaMA3-8B, Mistral-7B, and Phi-3.5-mini-instruct (3.8B). We agree that it would be interesting for future works exploring attribution or pruning strategies specific for MOE architectures.

#### 4. How an adversary might circumvent cache-level pruning

Thank you for raising this point. Conceptually, circumventing CachePrune is **challenging** because the defense is applied on the embedding-level (in KV Cache), thus does not rely on text clues that are visible to users, e.g., phrasing patterns or trigger words. In Figure 5 and Table 2, we show that CachePrune is robust to different types of attacks instructions and injected tasks. One hypothetical approach is to **compromise the software system and manually alter the learnt masking**, which could be out of the scope of machine learning threats.

#### 5. More detailed exploration of how pruning affects model internal representations, latency, and token probabilities

- **Influence on model internal representations:** In Figure 6, we have shown the **distribution of pruned neurons across layers**. As in Section 1, these neurons are associated with the LLM's view on **data vs. instruction** over the user-specified context. Ideally, we want the user-specified context to be treated as pure data. It can be observed that:
  - **The pruned neurons concentrates in the middle layers.** This aligns with previous works [6-8] showing that the middle layers are more capable of capturing abstract concepts, i.e., data or instruction.
  - There are generally **more key neurons being pruned than value neurons**. Since the key neurons controls the self-attention in Transformer, this suggests that **our approach works by intervening how a newly generated token attends to tokens from user-specified context**, so that it treats these tokens as data instead of instruction. In the meanwhile, the less pruning on the value shows that the pruning is preserving the encoded content (value) of the input context, thus maintaining the quality of clean responses that rely on this contextual knowledge.
- **Influence on latency:** The test-time overhead introduced by CachePrune is **negligible**, as it only involves masking (zeroing out) few neuron activations (e.g., 0.5%) in the KV cache.
- **Influence on token probabilities:** As mentioned in Section 2.4, the generation of clean vs. poisoned responses can be differed with only few triggering tokens. Our CachePrune performs attribution with logits of these triggering tokens (eq 13). Correspondly, the pruning **alters the generated token probabilities to favor tokens that trigger clean responses**.

[1] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

[2] LLaMA: Open and Efficient Foundation Language Models

[3] PaLM: Scaling Language Modeling with Pathways

[4] Sequence to Sequence Learning with Neural Networks

[5] [https://huggingface.co/docs/transformers/generation\\_strategies#greedy-search](https://huggingface.co/docs/transformers/generation_strategies#greedy-search)  
([https://huggingface.co/docs/transformers/generation\\_strategies#greedy-search](https://huggingface.co/docs/transformers/generation_strategies#greedy-search))

[6] Emergence of Abstractions: Concept Encoding and Decoding Mechanism for In-Context Learning in Transformers

[7] Does Representation Matter? Exploring Intermediate Layers in Large Language Models

[8] Learn when (not) to trust language models: A privacy-centric adaptive model-aware approach.




### Comparing to finetuning-based approach

Official Comment

by Authors (👁️ Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

📅 23 Nov 2025, 04:54 (modified: 17 Mar 2026, 09:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 1Lz7, Commitment Readers

 Revisions (/revisions?id=f4Q3wh4aoi)

#### Comment:

We train with DPO on the triplets of (prompt with injected instructions, poisoned response, clean response), encouraging the model to prefer clean responses over poisoned ones. For fair comparison, we train with sets of 8 triplets as in our experiments with CachePrune. Specifically, we sample 8 prompts with injected instructions, then sample the clean and poisoned responses as in Figure 7. The model is trained with LoRA following default parameters [9] in huggingface, except that we set the dropout rate as 0.05 (default 0.0) to mitigate overfitting when training on a small number of samples. We finetune LLama3-8b on SQuAD with batch size 4 for 20 epoches. In table below, we also include Datamarking as a strong prompt engineering baseline.

	ASR	F1 (clean)	F1 (attack)
Vanilla	27.86	28.20	19.56
Datamarking	13.25	28.56	21.45
CachePrune	<b>7.44 ± 0.22</b>	<b>28.68 ± 0.30</b>	<b>22.84 ± 0.49</b>
DPO (10 Epoches)	14.06 ± 2.50	27.48 ± 1.38	20.96 ± 2.10
DPO (20 Epoches)	8.05 ± 1.64	26.92 ± 1.26	19.71 ± 1.80

It can be observed that the funetuning yields comparable ASR to CachePrune while having slight degradation in F1. Comparatively, our CachePrune can be understood as a regularized funetuning:


- Instead of finetuning all the weights, it only modifies/prunes weights that induce the most salient features as in Section 2.2.1.
- To preserve the quality for clean responses, we further regularize by pruning only within a set  $\phi$  (Section 2.3).


[9] [https://huggingface.co/docs/peft/en/package\\_reference/lora](https://huggingface.co/docs/peft/en/package_reference/lora)  
([https://huggingface.co/docs/peft/en/package\\_reference/lora](https://huggingface.co/docs/peft/en/package_reference/lora))




## Gentle Follow-Up

Official Comment

by Authors ( Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

 24 Nov 2025, 19:44 (modified: 17 Mar 2026, 09:30)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer 1Lz7, Commitment Readers

 Revisions (/revisions?id=z9iM5Exb0g)

#### Comment:

As the author–reviewer discussion phase is approaching its end, we would kindly ask if you might have a chance to take a look at our rebuttal. We believe our responses have addressed the main concerns, and any additional feedback would be immensely helpful.

Thank you again for your time and thoughtful review.

## Official Review of Submission3650 by Reviewer qbZK

Official Review by Reviewer qbZK 📅 10 Nov 2025, 22:28 (modified: 17 Mar 2026, 09:30)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer qbZK, Authors, Commitment Readers

📄 Revisions (/revisions?id=cyTSzkwZVX)

### Paper Summary:

Introduces a novel defense for large language models that operates inside the model's KV-cache rather than at the prompt or training level. The method first identifies neurons whose activations trigger instruction-following behavior in the cached context. Using a preferential attribution loss derived from Direct Preference Optimization, it ranks neurons by their contribution to preferring poisoned versus clean responses. The top-scoring neurons are pruned from the KV cache, effectively forcing the model to treat the context purely as data.

### Summary Of Strengths:

The paper's strength lies in its clear and original rethinking of prompt-injection defense from a purely neural perspective. Instead of adding external rules or retraining the model, the authors look inside the KV cache to identify the neurons that cause an LLM to misinterpret contextual text as instructions and then prune them away. This approach is both elegant and practical—it requires no additional inference steps, integrates smoothly with existing systems, and preserves model performance. Empirically, CachePrune shows strong and consistent reductions in attack success rates across different models and datasets, with minimal loss in response quality. Moreover, the discovery of the “triggering effect,” where only a few early tokens determine whether the model is hijacked, offers a fresh mechanistic insight into how transformer activations govern behavior. Overall, the work combines interpretability, efficiency, and robustness in a way that feels both scientifically meaningful and operationally effective.

### Summary Of Weaknesses:

1. Requires the model or system to know where “context” ends and “instruction” begins—less realistic in free-form dialogue or retrieval settings.
2. Pruning neurons might suppress subtle reasoning behaviors or long-range dependencies if applied aggressively.
3. The preferential attribution loss depends on small sample sets; attribution noise could affect stability across runs.

### Comments Suggestions And Typos:

1. Could the authors provide a clearer analysis of what distinguishes the identified “instruction-triggering” neurons from others? For instance, are they concentrated in particular layers or attention heads, and are their activation patterns consistent across different prompts or models?
2. Beyond reducing attack success, does pruning these neurons affect the model's general reasoning or long-context understanding? A qualitative or quantitative comparison on benign tasks would strengthen confidence in safety-utility balance.

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Soundness:** 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.

**Excitement:** 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.

**Overall Assessment:** 4 = Conference: I think this paper could be accepted to an \*ACL conference.

### Ethical Concerns:

There are no concerns with this submission

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

**Publication Ethics Policy Compliance:** I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



## Reply to Reviewer qbZK

Official Comment

by Authors (👁️ Ryan A. Rossi (/profile?id=~Ryan\_A\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

📅 23 Nov 2025, 05:47 (modified: 17 Mar 2026, 09:30)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qbZK, Commitment Readers

📄 Revisions (/revisions?id=EeQZW1Ypjj)

### Comment:

Thank you so much for your thoughtful and constructive feedback!

#### 1. Requires the model or system to know where “context” ends and “instruction” begins.

We clarified this setup in Line 453-461. Specifically, this **aligns with LLM-integrated APIs** in which user queries are combined with third-party data using API-specific templates with pre-known positioning of the input context [1]. In addition, this is a **common setup also adopted in prior works [1-4]**.

#### 2. Pruning neurons might suppress subtle reasoning behaviors or long-range dependencies & Comparison on benign tasks

We add **reasoning experiments on benign tasks** that evaluate F1 (clean) on prompts without injection. Specifically, we allow the model generate a **Chain-of-Thought (COT)** before the final answer by appending the following to the prompt:

Before answering, first reason about the problem. In your final output, follow exactly this format (no extra text):

In the first line: Start the response with your reasoning. Do not add anything before.

From the second line and onward: Your final response and nothing else.

This makes the responses longer and tests the model's long-range dependency. Here are results on SQuAD with LLama3-8B with different pruning ratio  $p$  (default 0.5%). We also test with Datamarking which is the strongest baseline without reasoning.

	F1 (clean)
Vanilla + COT	39.56
Datamarking + COT	38.16
CachePrune + COT ( $p=0.5\%$ )	41.38 $\pm$ 0.79
CachePrune + COT ( $p=1.0\%$ )	38.27 $\pm$ 0.94

We can observe that CachePrune **does not affect reasoning with longer responses**, yielding comparable F1 (clean) with both  $p=0.5\%$  and  $p=1.0\%$ .

#### 3. Attribution loss depends on small sample sets & Attribution noise could affect stability

In Table 6, we show that CachePrune **works with larger sample sizes** by showing improved performance with more samples. Additionally, though the attribution noise can be a source of instability, we report in Table 1 that CachePrune yields **significantly lower ASR even considering the standard deviation**.

#### 4. A clearer analysis on the identified “instruction-triggering” neurons (concentrated in particular layers, etc.)

In Figure 6, we have shown the **distribution of pruned neurons across layers**. As in Section 1, these neurons are associated with the LLM's view on **data vs. instruction** over the user-specified context. Ideally, we want the user-specified context be treated as pure data. It can be observed that:

- **The pruned neurons concentrates in the middle layers**. This aligns with previous works [5-7] showing that the middle layers are more capable of capturing abstract concepts, i.e., data or instruction.
- There are generally **more key neurons being pruned than value neurons**. Since the key neurons controls the self-attention in Transformer, this suggests that **our approach works by intervening how a newly generated token attends to tokens from user-specified context**, so that it treats these tokens as data instead of instruction. In the meanwhile, the less pruning on the value shows that the pruning is preserving the encoded content (value) of the input context, thus maintaining the quality of clean responses that rely on this contextual knowledge.

[1] Jatmo: Prompt injection defense by task-specific finetuning.

[2] Struq: Defending against prompt injection with structured queries.

[3] Aligning llms to be robust against prompt injection.

[4] Are you still on track!? catching llm task drift with activations.

[5] Emergence of Abstractions: Concept Encoding and Decoding Mechanism for In-Context Learning in Transformers

[6] Does Representation Matter? Exploring Intermediate Layers in Large Language Models

[7] Learn when (not) to trust language models: A privacy-centric adaptive model-aware approach.

## Official Review of Submission3650 by Reviewer mfWA

Official Review by Reviewer mfWA 📅 10 Nov 2025, 18:23 (modified: 17 Mar 2026, 09:30)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Reviewer mfWA, Authors, Commitment Readers

📄 Revisions (/revisions?id=pC6YnqUjKI)

### Paper Summary:

This paper introduces a neural-based attribution method for defending against prompt injection. The approach first identifies neurons that contribute significantly more to instruction following than to context processing, and then intervenes by suppressing these neurons within the context representations. Experiments demonstrate that the proposed method outperforms baselines in defending against prompt injection attacks.

### Adequacy Of Revisions:

The authors have adequately addressed the concerns in previous reviews.

### Summary Of Strengths:

The proposed method is lightweight and requires no heavy training, yet it demonstrates clear improvements over prompt-based baselines. The approach is well-motivated, and its effectiveness is validated by the experimental results.

### Summary Of Weaknesses:

My primary concern is that the baselines are all prompt-based methods that apply to both white and black-box models. Since this method requires modification in the parameter space and only applies to white-box LLMs, I think it's more fair to also compare to finetuning methods for prompt injection defense, such as SFT or DPO that discourages the model to follow the injected prompts.

### Comments Suggestions And Typos:

Most equations in the paper are missing punctuation at the end.

**Confidence:** 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

**Soundness:** 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.

**Excitement:** 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.

**Overall Assessment:** 3 = Findings: I think this paper could be accepted to the Findings of the ACL.

**Ethical Concerns:**

There are no concerns with this submission

**Needs Ethics Review:** No

**Reproducibility:** 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

**Datasets:** 1 = No usable datasets submitted.

**Software:** 1 = No usable software released.

**Knowledge Of Or Educated Guess At Author Identity:** No

**Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Knowledge Of Paper Source:** N/A, I do not know anything about the paper from outside sources

**Impact Of Knowledge Of Paper:** N/A, I do not know anything about the paper from outside sources

**Reviewer Certification:** I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

**Publication Ethics Policy Compliance:** I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



## Reply to Reviewer mfwA

Official Comment

by Authors ( Ryan A. Rossi (/profile?id=~Ryan\_A.\_Rossi2), Rui Wang (/profile?id=~Rui\_Wang25), Ruiyi Zhang (/profile?id=~Ruiyi\_Zhang3), Yu Xia (/profile?id=~Yu\_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

23 Nov 2025, 04:14 (modified: 17 Mar 2026, 09:30)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer mfwA, Commitment Readers

Revisions (/revisions?id=1j3dF1x2KW)

**Comment:**

Thank you so much for your constructive and insightful suggestions!

We train with DPO on the triplets of (prompt with injected instructions, poisoned response, clean response), encouraging the model to prefer clean responses over poisoned ones. For fair comparison, we train with sets of 8 triplets as in our experiments with CachePrune. Specifically, we sample 8 prompts with injected instructions, then sample the clean and poisoned responses as in Figure 7. The model is trained with LoRA following default parameters [1] in huggingface, except that we set the dropout rate as 0.05 (default 0.0) to mitigate overfitting when training on a small number of samples. We finetune LLama3-8b on SQuAD with batch size 4 for 20 epoches. In table below, we also include Datamarking as a strong prompt engineering baseline.

	ASR	F1 (clean)	F1 (attack)
Vanilla	27.86	28.20	19.56
Datamarking	13.25	28.56	21.45
CachePrune	<b>7.44 ± 0.22</b>	<b>28.68 ± 0.30</b>	<b>22.84 ± 0.49</b>
DPO (10 Epoches)	14.06 ± 2.50	27.48 ± 1.38	20.96 ± 2.10
DPO (20 Epoches)	8.05 ± 1.64	26.92 ± 1.26	19.71 ± 1.80

It can be observed that the funetuning yields comparable ASR to CachePrune while having slight degradation in F1. Comparatively, our CachePrune can be understood as a regularized funetuning:

- Instead of finetuning all the weights, it only modifies/prunes weights that induce the most salient features as in Section 2.2.1.
- To preserve the quality for clean responses, we further regularize by pruning only within a set  $\phi$  (Section 2.3).

[1] [https://huggingface.co/docs/peft/en/package\\_reference/lora](https://huggingface.co/docs/peft/en/package_reference/lora)  
([https://huggingface.co/docs/peft/en/package\\_reference/lora](https://huggingface.co/docs/peft/en/package_reference/lora))



## Gentle Follow-Up

Official Comment

by Authors (👁️ [Ryan A. Rossi \(/profile?id=~Ryan\\_A.\\_Rossi2\)](/profile?id=~Ryan_A._Rossi2), [Rui Wang \(/profile?id=~Rui\\_Wang25\)](/profile?id=~Rui_Wang25), [Ruiyi Zhang \(/profile?id=~Ruiyi\\_Zhang3\)](/profile?id=~Ruiyi_Zhang3), [Yu Xia \(/profile?id=~Yu\\_Xia9\)](/profile?id=~Yu_Xia9), +5 more (/group/edit?id=aclweb.org/ACL/ARR/2025/October/Submission3650/Authors))

📅 24 Nov 2025, 19:46 (modified: 17 Mar 2026, 09:30)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer mfWA, Commitment Readers

📄 Revisions (/revisions?id=Os8F6JtDtb)

### Comment:

As the author-reviewer discussion phase is approaching its end, we would kindly ask if you might have a chance to take a look at our rebuttal. We believe our responses have addressed the main concern, and any additional feedback would be immensely helpful.

Thank you again for your engagement in the reviewing process.

FAQ (<https://docs.openreview.net/getting-started/frequently-asked-questions>)

[id=OpenReview.net/Support](https://openreview.net/Support)

[Contact \(/contact\)](/contact)

[All Venues \(/venues\)](/venues)

[Donate \(/donate\)](/donate)

[Sponsors \(/sponsors\)](/sponsors)

[Terms of Use \(/legal/terms\)](/legal/terms)

[News \(/group?id=OpenReview.net/News&referrer=\[Homepage\]\(\)\)](/group?id=OpenReview.net/News&referrer=[Homepage]())

[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors). © 2026 OpenReview