# ParamMute: Suppressing Knowledge-Critical FFNs for Faithful Retrieval-Augmented Generation

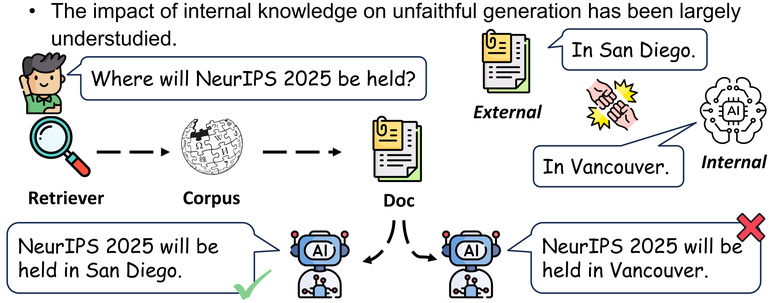
Pengcheng Huang[1], Zhenghao Liu[1], Yukun Yan[2], Haiyan Zhao[2], Xiaoyuan Yi[3], Hao Chen[2], Zhiyuan Liu[2], Maosong Sun[2], Tong Xiao[1], Ge Yu[1], Chenyan Xiong[4]

1 NEU-China 2 THU-NLP 3 MSRA 4 CMU

NEURAL INFORMATION PROCESSING SYSTEMS

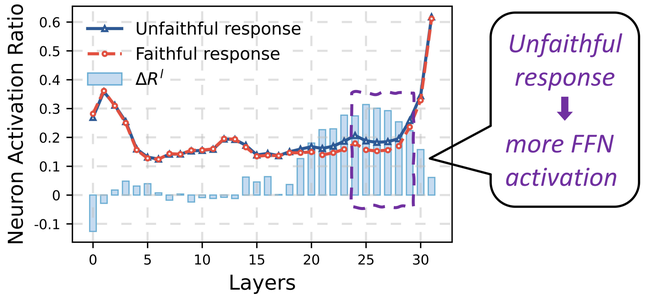
## Motivation and Background

- **Retrieval-Augmented Generation**
  - RAG models retrieve documents from the external corpus and then augment the LLM's generation.
  - They help LLMs alleviate hallucinations and generate more accurate and trustworthy responses.
- **Contextual Faithfulness in Retrieval-Augmented Generation**
  - RAG models can generate outputs that are irrelevant to or even contradict the retrieved content.
  - Prior work has mainly concentrated on improving the utilization of external knowledge in RAG models.
  - The impact of internal knowledge on unfaithful generation has been largely understudied.




## Preliminaries

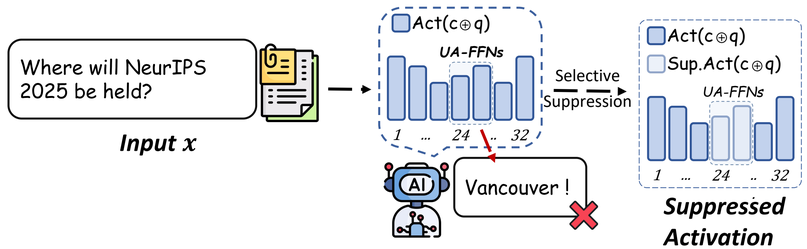
- **Understanding the Role of FFN in Unfaithful Generation**
  - FFNs function similarly to **key-value memory** mechanisms, storing the majority of LLMs' parametric knowledge.
  - $FFN(x_i^l) = (\sigma(K^l \cdot x_i^l))^T V^l = \sum_{j=1}^{d_m} \sigma(x_i^l \cdot k_j^l) v_j^l = \sum_{j=1}^{d_m} a_{ij}^l v_j^l$
- **We use FFN activation ratio as a proxy to study how parametric knowledge affects unfaithful generation.**
  - **Unfaithful response:** output that is irrelevant to or contradicts the retrieved content.
  - **Faithful response:** output that aligns with the retrieved context.



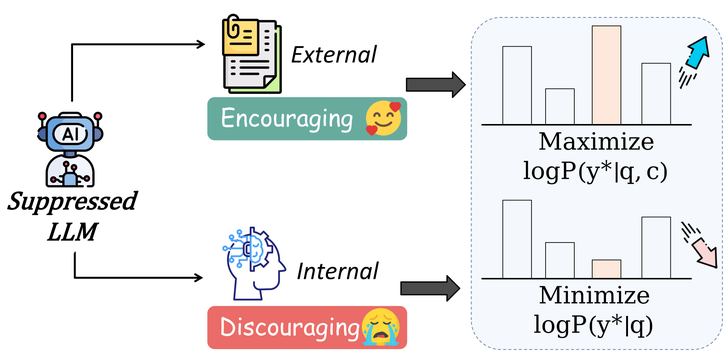

- When **mid-to-deep FFN** ***(Unfaithfulness-Associated FFNs, UA-FFNs)*** layers exhibit excessive activation, the model tends to rely more heavily on its internal knowledge, consequently leading to unfaithful outputs.

## Parametric Knowledge Muting through FFN Suppression (ParamMute)

- **Activation Suppression (§3.1)**
  - To reduce the model's reliance on parametric knowledge, we suppress the activations of Unfaithfulness-Associated FFNs.

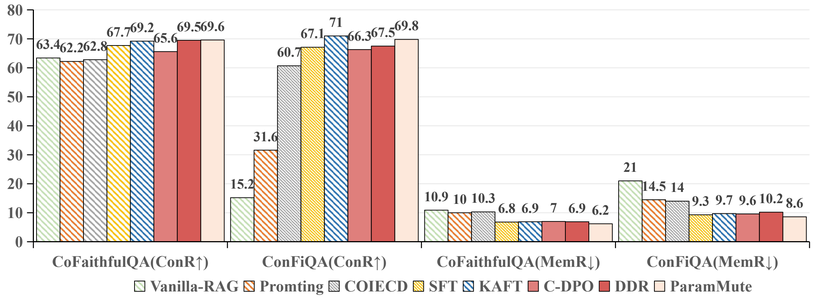



- **Knowledge-Augmented Adaptation (§3.2)**
  - To further recalibrate the model's knowledge utilization preferences, we incorporate a plug-and-play adaptation module that enables more effective use of external evidence.



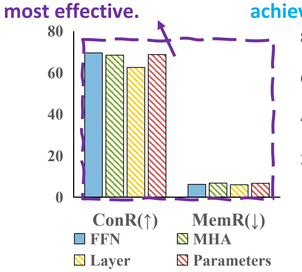

## Experimental Results

- **Overall Performance**
  - ParamMute outperforms all baseline models across datasets.

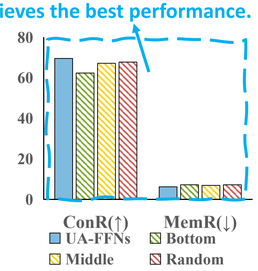



- **Ablation and Component Analysis**

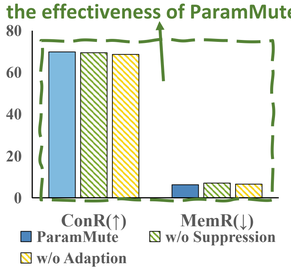
Suppressing FFNs is most effective.



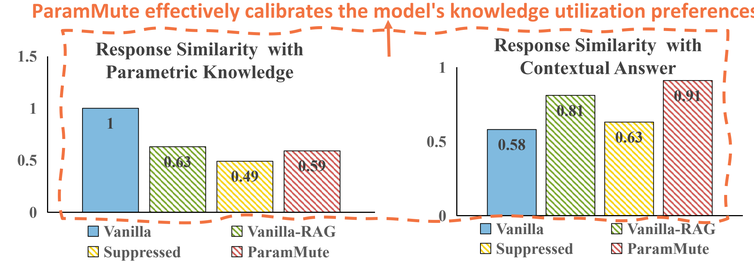

Suppressing the UA-FFNs achieves the best performance.



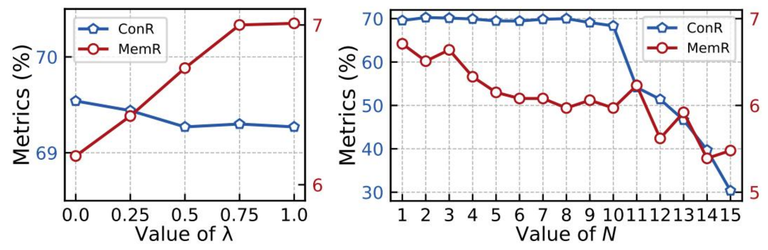

Both components contribute to the effectiveness of ParamMute.




ParamMute effectively calibrates the model's knowledge utilization preferences.




- **Parameter Sensitivity Analysis**




- Smaller values of λ consistently improve contextual faithfulness while reducing the model's dependence on internal memory.
- Increasing N enlarges the suppression scope and steadily decreases MemR; however, when N exceeds 10, ConR drops sharply, indicating that excessive suppression may disrupt essential functions beyond knowledge storage.

## Resources

Paper

Code

Wechat

Explore more interesting experiments in our paper and code, and feel free to contact us at pengcheng.neu@outlook.com