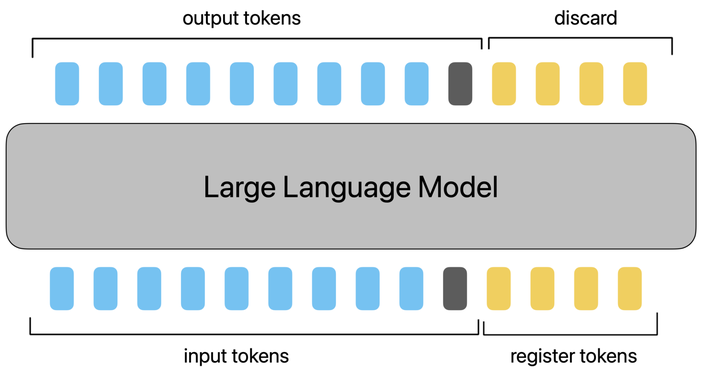
Register-Augmented LLM Fine-Tuning

ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

Era Parihar, Ishan Kapnadak, Nilay Gautam, Rishikesh Ksheersagar, Shlok Agarwal
{erap, kapnadak, gnilay, rishiksh, ashlok}@umich.edu
University of Michigan, Ann Arbor

Introduction

- Pre-trained LLMs have shown strong performance on various NLP tasks, including Question-Answering (QA)
- However, fine-tuning for QA exposes inefficiencies in managing high-norm tokens (artefacts) leading to suboptimal performance
- Recent developments in Vision Transformers have advocated for adding register tokens to transformers – allowing the model to focus on task-relevant details by better managing global context
- We propose a novel application of register-augmentation in language models to mitigate the impact of artefacts and enhance performance, learning efficiency, and interpretability



Attention Analysis

- We use two methods for analyzing and interpreting attention heads: integrated gradients and layer-wise relevance propagation (LRP)

Layer-wise Relevance Propagation

- LRP computes relevance scores starting from the output layer and propagates them backwards all the way to the input layer
- Once the propagation is done, each input token's relevance score indicates how much it contributed to the model output
- This allows us to track evolution of attention heads and allows for interpretability and transparency of the model

Integrated Gradients

- Integrated Gradients were used to conduct attribution analysis of QA model to quantify the influence of each token in model output
- Attribution scores were overlayed onto the sequence tokens, with green indicating positive contribution and red indicating negative
- This allows us to track which tokens influence our model output, both positively and negatively, and which don't contribute as much

Implementation

- A BERT Encoder model is used as the baseline for our experiments
- RegBERT inherits from BERT and prepends register tokens to the input sequence during the forward pass and are removed before predicting the start and end tokens
- We also create RegBERTForQA using BERTForQA which uses the RegBERT class as its BERT Model
- We analyze our results using F1 scores and ExactMatch, as well as attention map analysis through LRP and Integrated Gradients. These are compared for BERT with & without register tokens

Experimental Results

Integrated gradients- Without registers


Visualizations For Start Position

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
15	15 (0.43)	13	2.16	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs . it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]

Visualizations For End Position

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
29	29 (0.66)	14	3.13	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs , it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]

Integrated gradients- With registers


Visualizations For Start Position

Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
21	21 (0.46)	13	3.11	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs , it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]

Visualizations For End Position

Legend: Negative Neutral Positive

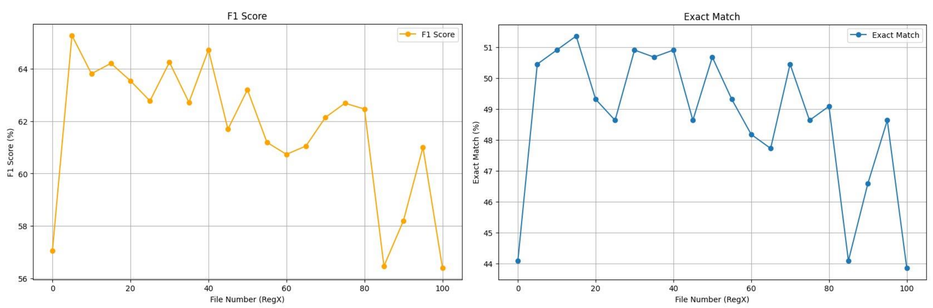
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
29	29 (0.80)	14	3.02	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs , it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]

Without registers

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
encourages and speeds wound healing via cleaning and protection from reinjury or infection (0.56)			1.00	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs , it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]

With registers

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
wound care encourages and speeds wound healing via cleaning and protection from reinjury or infection (0.01)			1.00	[REG0] [REG1] [REG2] [REG3] [REG4] [REG5] [REG6] [REG7] [REG8] [REG9] [REG10] [REG11] [REG12] [REG13] [REG14] [REG15] [REG16] [REG17] [REG18] [REG19] [REG20] [REG21] [REG22] [REG23] [REG24] [REG25] [REG26] [REG27] [REG28] [REG29] [REG30] [REG31] [REG32] [REG33] [REG34] [REG35] [REG36] [REG37] [REG38] [REG39] [REG40] [REG41] [REG42] [REG43] [REG44] [REG45] [REG46] [REG47] [REG48] [REG49] [CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' s needs , it can range from the simplest first aid to entire nursing special ##ties such as wound , os ##tom ##y , and con ##tine ##nce nursing and burn center care . [SEP]



Experimental Results

- LRP shows better and more informative analysis of the attention map as compared to the Integrated Gradients approach
- On analyzing the attention maps, we see that the model fine-tuned with registers has significantly less noise in its attention map
- Even if the register-augmented model cannot predict the exact answer, it comes up with a better answer then the regular model
- Finally, both F1 score and ExactMatch show a significant improvement from the regular model (num_registers = 0) when augmented with registers (num_registers > 0)
- However, this increase dies down as we increase the number of registers, indicating that there is a sweet spot in the middle where register-augmentation is most effective at improving performance

Conclusion and Future work

- Overall, as seen before in vision transformers, augmenting the fine-tuning procedure with registers improves performance of LLMs
- This is demonstrated by an increase in both F1 score and ExactMatch as well as improved interpretability on adding registers
- However, our scope of study is limited to only one Transformer model and one task, which does not allow us to test robustness
- Finally, we plan to extend our study to other tasks and domains in the future, with the following being interesting avenues to explore:
  a. Register-Augmentation in Multilingual/Cross-lingual settings: Does register-augmentation work for low-resource languages? Does the number of registers vary for different languages?
  b. Perform more-fine grained attention analysis: How does attention distribute across different registers? Does each register play an equal role in augmentation?
  c. Generalize to more NLP tasks apart from QA: Does the augmented model still outperform regular model?
  d. Combine register-augmentation with other techniques and analyze if registers work with perturbations and noise in input

Significant References

- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024.
- Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021.
- Mikhail S. Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V. Sapunov. Memory transformer, 2021.

COLLEGE OF ENGINEERING
ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN