EECS 598 (LLMs): Project Report
Register-Augmented LLM Fine-Tuning

Era Parihar erap@umich.edu
Ishan Kapnadak kapnadak@umich.edu
Nilay Gautam gnilay@umich.edu
Rishikesh Ksheersagar rishiksh@umich.edu
Shlok Agarwal ashlok@umich.edu

Abstract

Pre-trained LLMs have shown strong performance across various NLP tasks including Question Answering (QA). Despite their strong generalizability and exceptional adaptability, LLMs suffer from long-range dependencies and artifacts when fine-tuned for specific tasks, leading to suboptimal performance. Recent developments in Vision Transformers have advocated for the use of some auxiliary register tokens during the fine-tuning procedure to allow the model to focus on task-relevant details by better management of global information, leading to provably better interpretability and performance. We propose and study a novel application of register-augmented fine-tuning for LLMs in the language domain, and show that register-augmented language models can better mitigate the effects of artifacts, leading to higher interpretability, less noise, and overall better performance.

1 Introduction

Pre-trained large language models (LLMs), such as BERT and LLaMA, exhibit strong performance across various NLP tasks, but fine-tuning for specific tasks like question answering (QA) often exposes inefficiencies in managing long-range dependencies and artifacts in attention mechanisms. These artifacts, typically high-norm tokens, lead to suboptimal performance by drawing excessive attention to less important parts of the input.

Recent studies have investigated the use of register tokens in transformer models. Specifically, in Vision Transformers, augmenting the input sequence with empty register tokens has shown to address artifacts and high-norm tokens, by offloading information from high-norm tokens and allowing the model to focus on task-relevant details, and to much better capture global context.

Our project introduces register tokens to language models to manage global information more effectively. By outsourcing essential computations to these tokens, register-augmented LLMs mitigate the impact of artifacts and high-norm tokens in their attention mechanisms, resulting in more focused and interpretable attention maps, enhancing performance, learning efficiency, as well as interpretability.

2 Related Work

To better understand the functionality and the performance of a Large Language Model it is imperative that we first understand the key part of the model that is the attention map. Recent research has made notable strides in interpreting the attention mechanisms within Vision Transformers (ViTs). For example, Dosovitskiy et al. [2021] Darcet et al. [2024] have provided insights into visualizing and interpreting self-attention in vision models. Expanding on this, Chefer et al. [2021] introduced a

novel approach for calculating relevancy in transformer networks using a deep Taylor decomposition framework Chefer et al. [2021]. This method computes local relevancy scores and propagates them through the attention layers, while maintaining the total relevancy across layers. Chefer's approach has been tested on both vision and language tasks, demonstrating its ability to identify salient patches in images that influence classification decisions. In the case of language models, the method successfully extracts input components that support correct classification, as demonstrated using a fine-tuned BERT model on the Movie Reviews dataset.

In addition to the relevancy-based methods, there are approaches specifically designed to enhance the performance of Vision Transformers. One such method proposes adding additional dummy tokens to the input sequence to improve the classification accuracy of ViTs Darcet et al. [2024]. The authors observed that tokens with high norms in attention maps often carry significant information for a given class, but are sometimes concentrated in low-information regions of the image. By introducing extra register tokens, they aim to redirect the information from these high-norm tokens to the newly added tokens, thereby improving the models classification performance.

In our work we take inspiration from the vision domain and use the additional register token to learn the global context of a given task and improve performance of the fine tuned model. Darcet et al. [2024] shows that Norms of the tokens in the areas of the image where the background of the image has very low information (That is there are no high frequency color changes/minimum gradient changes and etc.). Whereas in the model with registers appended before the sequence it is observed that the number of patches with higher norms in the background area of the image decrease significantly. From this observation it was concluded that with registers the model focuses more on the actual subject in the image. The registers help by offloading some internal computation to reduce interference with local information processing, some of these high-norm tokens remain essential for capturing global patterns across the image. Similarly in the Language context these additional tokens can learn the global context of the task that the model is being fine-tuned for and in turn can enhance the performance of the model.

3 Methodology

3.1 Register-Augmented Fine-Tuning

Our principle approach is to add register tokens to the transformer model during fine-tuning, similar to the Computer Vision approach proposed in Darcet et al. [2024]. We define a register-augmented BERT model that uses BERT's underlying architecture but appends empty register tokens to the input sequence during fine-tuning. Concretely, if our model is fed in the input sequence $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{N \times d}$, then our register-augmented model manually constructs the modified input sequence $\hat{\mathbf{x}} = [\mathbf{r}_1, \ldots, \mathbf{r}_\tau, \mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{(N+\tau) \times d}$ by appending τ empty register tokens at the front. The model is then fine-tuned on these augmented sequences. Our motive is to allow the transformer model to use these empty register tokens to extract and store useful global context information from the actual sequence $\mathbf{x}$ to enhance our performance on the task at hand. Schematically, our model is represented in Figure 1

3.2 Attention Analysis

Along with increased performance, we also wish to analyze if augmenting with registers actually helps the transformer to make use of global information better. To do so, we resort to performing attention analysis to see how the interpretability of the transformer changes, by analyzing the contribution of various tokens to the final model output. We use two methods for attention analysis – (i) layer-wise relevance propagation and (ii) integrated gradients.

3.2.1 Layer-wise Relevance Propagation

Layer-wise relevance propagation, as proposed by Chefer et al. [2021], is a method to compute relevancy scores for each token based on how much they contribute to the model output. This method uses the Deep Taylor Decomposition principle to compute local relevance scores for the output and propagates these scores backwards through the model to obtain relevance scores at the input end using an approach similar to backpropagation. More concretely, let us label our layers as $L^{(n)}(\mathbf{X}, \mathbf{Y})$

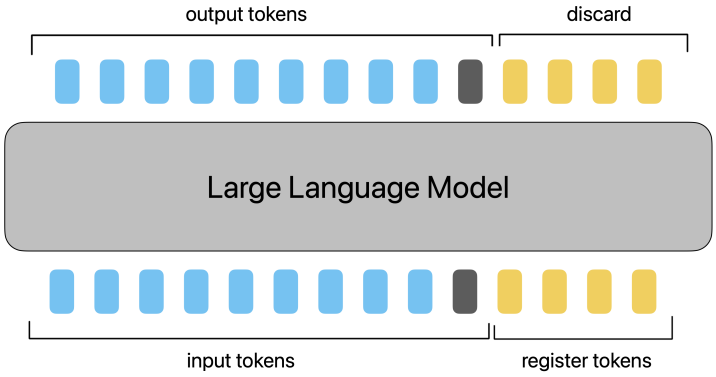


Figure 1: Schematic of Register-Augmented LLMs (Adapted from Darcet et al. [2024])

denote the output of the n^{th} layer on two tensors $\mathbf{X}$ and $\mathbf{Y}$ (usually the input features and the weights). With this, we use the following generic Deep Taylor Decomposition (as per Montavon et al. [2017])

$$R_j^{(n)} = \mathcal{G}(\mathbf{X}, \mathbf{Y}, R^{(n-1)}) = \sum_i \mathbf{X}_j \frac{\partial L_i^{(n)}(\mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}_j} \frac{R_i^{(n-1)}}{L_i^{(n)}(\mathbf{X}, \mathbf{Y})}.$$

where $R^{(n)}, R^{(n-1)}$ denote the relevance scores at the n^{th} and $(n-1)^{\text{th}}$ layers respectively. Note that since we are backpropagating relevance scores, layer 1 is the output layer.

3.2.2 Integrated Gradients

Whereas LRP takes a computational approach to compute relevance scores, Sundararajan et al. [2017] propose an axiomatic approach to define attribution scores – that is, an axiomatic approach to quantify how much each input token contributes to the model output. For a function F implemented by a deep network and input $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, one can axiomatically define an attribution score (relative to some baseline input $\mathbf{x}'$)) $A_F(\mathbf{x}, \mathbf{x}') = (a_1, \ldots, a_n) \in \mathbb{R}^n$ where a_i captures the contribution of x_i in the prediction $F(\mathbf{x})$.

An attribution score is constructed with the use of two fundamental and desired axiom – (i) sensitivity and (ii) implementation invariance. Intuitively, sensitivity says that if the input and baseline differ in only one feature and have different predictions, then the differing feature should be given a non-zero attribution score. Implementation invariances says that if two networks implement the same function (i.e., the two networks are functionally equivalent), then the resulting attribution scores should be the same.

Sundararajan et al. [2017] propose the integrated gradients scheme, defined as follows:

$$\mathsf{IntegratedGrads}_i(\mathbf{x}) := (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial \mathbf{x}_i} \, d\alpha.$$

Integrated Gradients satisfy the two desired axioms, and in fact satisfy a much stronger axiom of completeness. That is, if $F\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable almost everywhere, then

$$\sum_{i=1}^{n} \mathsf{IntegratedGrads}_i = F(\mathbf{x}) - F(\mathbf{x}').$$

Further, by picking our baseline such that $F(\mathbf{x}') \approx 0$, integrated gradients give an easy feature-wise decomposition of the model prediction $F(\mathbf{x})$.

3.3 Implementation Details

Two key components form the backbone of this architecture: `RegBert` and `RegBertForQA`. The `RegBert` class extends the standard BERT model to incorporate register tokens and their corresponding positional embeddings. These tokens are appended to the input sequence embeddings, and the

attention mask is modified to include them. During the forward pass, the combined embeddings are processed by the encoder, ensuring that the register tokens participate in the model's attention and learning mechanisms. Additionally, the `RegBert` class supports relevance propagation, enabling analysis of how information flows through the model layers and identifying the contributions of the register tokens.

The `RegBertForQA` class builds on `RegBert`, specifically adapting it for Question-Answering tasks. It outputs start and end logits for answer span predictions while excluding register tokens during inference. The QA-specific adjustments include modifications to the final linear layer and loss computation to ensure compatibility with the extended sequence length. This class also integrates relevance propagation techniques, facilitating interpretability by tracing how register tokens influence the models predictions.

The implementation is tailored for Question-Answering (QA) tasks, where the model processes preprocessed sequences using a maximum length of 434 tokens and a stride of 128. Training is conducted with a learning rate of 2×10^{-5}, three epochs, and weight decay of 0.01. During inference, the outputs are adjusted to exclude the register tokens from predictions, ensuring accuracy in the original input context. The English corpus from TyDi QA GoldP dataset serves as the basis for both training and evaluation.

4 Experimental Results

Our experimental evaluation focuses on two key aspects: (1) the quantitative performance improvements achieved through register augmentation, and (2) the qualitative analysis of attention patterns using both Layer-wise Relevance Propagation (LRP) and Integrated Gradients methods.

4.1 Performance Analysis

We conducted an extensive ablation study to determine the optimal number of register tokens for our task. Figure 4a presents the results of this analysis using two primary metrics: F1 score and Exact Match (EM) score. Several key findings emerge from this analysis:

- The addition of register tokens consistently improves model performance up to a certain point. Both F1 and EM scores show significant improvement when moving from 0 registers (baseline) to 5–10 registers.
- There exists an optimal range for the number of registers, beyond which performance begins to degrade. This suggests that while register augmentation is beneficial, there is a sweet spot that balances the model's ability to capture global context without overwhelming the attention mechanism.
- The performance curves for both F1 and EM scores follow similar patterns, indicating that the benefits of register augmentation are consistent across different evaluation metrics.

4.2 Attention Analysis

To better understand how register tokens influence the model's behavior, we employed two complementary analysis techniques: Layer-wise Relevance Propagation (LRP) and Integrated Gradients. The visualizations in Figures 3b and 4b reveal several interesting patterns:

4.2.1 Layer-wise Relevance Propagation Analysis

The LRP analysis (Figure 3b) demonstrates notable differences between the standard and register-augmented models:

- Without registers (Figure 4a), the attention patterns appear more diffused, with relevance scores spread across multiple tokens.
- With registers (Figure 3b), the attention becomes more focused, suggesting that the register tokens help the model better identify and utilize relevant information in the input sequence.

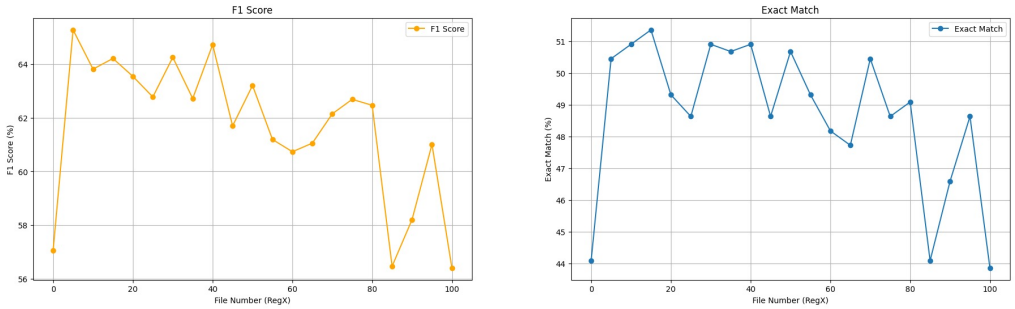
4.2.2 Integrated Gradients Analysis

The Integrated Gradients visualization (Figure 4b) provides additional insights into the token attribution patterns:

- The baseline model (Figure 4a) shows less structured attribution patterns, potentially indicating less efficient information processing.
- The register-augmented model (Figure 4b) exhibits more organized attribution patterns, suggesting that the registers help the model develop a more systematic approach to processing the input sequence.

These qualitative analyses support our quantitative findings, indicating that register tokens not only improve performance metrics but also enhance the model's ability to process and utilize information more effectively. The visualizations suggest that registers help the model better manage global context while maintaining focus on relevant local information.

The combined results from both quantitative and qualitative analyses provide strong evidence that register augmentation is an effective approach for improving transformer-based language models in the question-answering task. The improvements in both performance metrics and attention patterns indicate that the benefits of register augmentation extend beyond simple performance gains to fundamental improvements in how the model processes and utilizes information.



(a) F1 Score v/s Number of Registers Augmented (b) Exact Match v/s Number of Registers Augmented

Figure 2: Ablation study for number of registers with F1 score and exact match as metrics

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]

(a) Layer-wise Relevance Propagation without 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 (0.61)		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]

(b) Layer-wise Relevance Propagation with registers

Figure 3: LRP Visualization with and without registers

5 Future Work & Conclusions

From our experimental results, we draw the conclusion that register-augmentation in the fine-tuning procedure aids LLMs to better manage global context and improve performance. This is demonstrated

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]

(a) Integrated Gradients without 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 (0.85)	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]

(b) Integrated Gradients with registers

Figure 4: Integrated Gradients Visualization with and without registers

by a stark increase in F1 score and Exact Match when the number of registers used increases from 0 (regular fine-tuning) to 5 or 10 (register-augmented fine-tuning). Although the performance decreases on adding too many registers, there is a sweet spot where register augmentation significantly aids LLM performance.

However, our study is limited to only one Transformer model (BERT) and one NLP task (Question Answering). As such, we have not tested for the robustness of our method against various architectures and tasks. With this in mind, we have some future directions to expand our work, and some questions that we would like to investigate.

1. **Register Augmentation in Multilingual Settings.** Does register augmentation work for low-resource languages? Does the number of registers required for significant increase in performance vary across different languages?
2. **Attention Analysis.** We would like to investigate how the attention actually distributes across each register, and ultimately understand if all registers play an equal role in our model behaviour, rather than some registers having a greater influence than others.
3. **Generalizability.** Generalize to other NLP tasks besides QA and to other Transformer models besides BERT. Does the register-augmented model still outperform the regular one?

Author Contributions

The work for this project was equitably divided between all members. Era and Nilay were involved in coding up and implementing the attention analysis including LRP and Integrated Gradients. Shlok and Rishikesh were involved in coding up the register-augmented BERT model and performing

ablation studies. Ishan was involved in debugging and training the models and attention maps, and preparing the poster. The final report had contributions from all group members, with Ishan writing the abstract, introduction and methodology, Era and Nilay writing the experimental results section, Shlok writing the related works section, and Rishikesh writing the implementation details.

References

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization, 2021. URL https://arxiv.org/abs/2012.09838.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2024. URL https://arxiv.org/abs/2309.16588.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2016.11.008. URL https://www.sciencedirect.com/science/article/pii/S0031320316303582.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017. URL https://arxiv.org/abs/1703.01365.

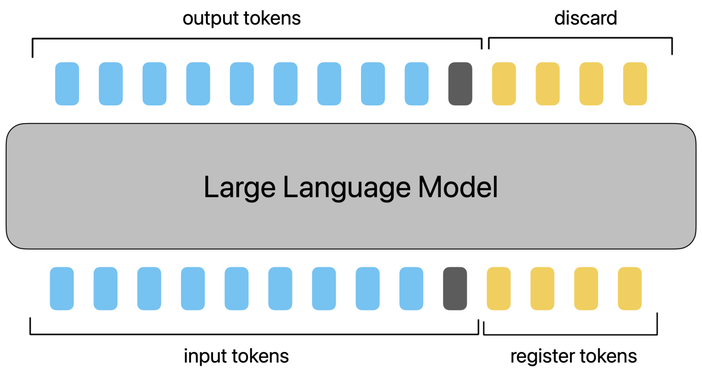
Register-Augmented LLM Fine-Tuning

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

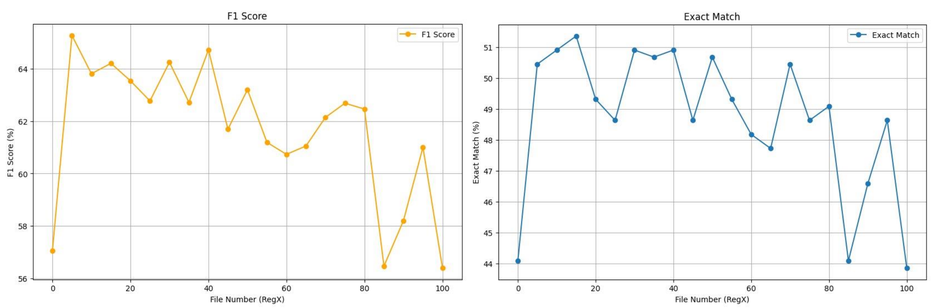
True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
29	29 (0.80)	14	3.02	[CLS] what is a way to increase your wound healing speed ? [SEP] wound care encourages and speeds wound healing via cleaning and protection from rein ##ju ##ry or infection , depending on each patient ' 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