Interpreting Multilingual and Document-Length Sensitive Relevance Computations in Neural Retrieval Models through Axiomatic Causal Interventions

Oliver Savolainen[*]
University of Amsterdam
Amsterdam, Netherlands
oliver.savolainen@student.uva.nl

Dur e Najaf Amjad[*]
University of Amsterdam
Amsterdam, Netherlands
najaf.amjad@student.uva.nl

Roxana Petcu
University of Amsterdam
Amsterdam, Netherlands
r.m.petcu@uva.nl

Abstract

This reproducibility study analyzes and extends the paper "Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models," which investigates how neural retrieval models encode task-relevant properties such as term frequency. We reproduce key experiments from the original paper, confirming that information on query terms is captured in the model encoding. We extend this work by applying activation patching to Spanish and Chinese datasets and by exploring whether document-length information is encoded in the model as well. Our results confirm that the designed activation patching method can isolate the behavior to specific components and tokens in neural retrieval models. Moreover, our findings indicate that the location of term frequency generalizes across languages and that in later layers, the information for sequence-level tasks is represented in the CLS token. The results highlight the need for further research into interpretability in information retrieval and reproducibility in machine learning research. Our code is available at https://github.com/OliverSavolainen/axiomatic-ir-reproduce.

CCS Concepts

• **Information systems → Retrieval models and ranking**.

Keywords

interpretability; neural ranking models; information retrieval axioms; multilingual retrieval

1 Introduction

Transformers have achieved state-of-the-art performance across various domains, including information retrieval. Their inner workings have been extensively studied, to make their reasoning process

[*]These authors contributed equally to the paper.

more transparent and trustworthy. The paper "Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models" [4] explores an approach to understanding how a neural retrieval model makes its modelling decisions.

Original findings. Chen et al. [4] apply activation patching to TAS-B [10], a BERT-based neural retrieval model. The activation patching method uses a carefully crafted diagnostic dataset that tries to isolate term frequency information through synthetic query-document pairs [7]. This work presents the following key findings:

1) Activation patching can effectively isolate a task-dependent behaviour to specific components and tokens in neural retrieval models.

2) Term frequency information is stored across the occurrences of the query term, with a focus on the first term positions in the document, in the first layers of the model. Later, this information is stored in the pooled representation of the CLS token.

3) The attention mechanism includes four attention heads specific for processing term frequency information. These heads work as two different purpose groups in the initial layers and middle layers that interact with each other. [4]

Our findings. To validate the claims made, we aim to reproduce and further extend the paper by Chen et al. [4]. The first extension examines the inner workings of the model across different languages, specifically Spanish and Chinese. The other extension uses the LNC1 axiom to investigate whether the model encodes information on document length, a critical feature for effective ranking models. Our experiments confirm that the model stores term frequency information across different languages, with an emphasis on the first positions of the document. We also show that the model stores document length information at the CLS token. While we think that the designed method can isolate the behaviour of specific components of neural retrieval models, it could be backed up by other interpretability methods such as circuit analysis and sparse autoencoders [6, 12].

The paper is organized as follows. In Section 2, we review the relevant literature. Section 3 outlines the methodology used in the original study. In Section 4, we introduce two extensions to the original approach. In Section 5, we describe the experimental setup and in Section 6 the results of both the reproduced and the extension experiments are discussed. Section 7 consists of discussing the original paper and findings from our work. Finally, conclusions are made in Section 8.

2 Related Work

Activation Patching. Causal interventions differentiate themselves from other interpretability methods, such as probing [1,

5], by providing causal insights into the model behaviour. One such method is activation patching, for which a corrupted input $X_{\text{perturbed}}$ with its output $Y_{\text{perturbed}}$ is needed alongside a clean input X_{baseline} with the correct output Y_{baseline}. For example, X_{baseline} could be "Paris is the capital of", with Y_{baseline} "France" while $X_{\text{perturbed}}$ is "London is the capital of" and the expected answer Y_{corrupt} for it is "England." To perform activation patching, the model is first run with X_{baseline}, and its activations are cached. The model is then run with $X_{\text{perturbed}}$, and selected activations are replaced with those from the run with X_{baseline}. To understand the impact of specific components from the model, we measure how well we can recreate the result of X_{baseline} by iteratively replacing activations. In the example, it would be about how close we reach to getting Y_{baseline} "France" [4, 14, 16].

TFC1 and LNC1 Axioms. Axiomatic Information Retrieval (IR) has focused on defining retrieval principles to guide neural ranking models and their evaluation. The TFC1 axiom [7] examines the relationship between term frequency and retrieval effectiveness, emphasizing that higher query term frequency in a document should lead to higher document relevance scores, provided all other factors remain equal. The LNC1 axiom [8] states that if two documents contain the same terms with the same term frequencies, but one document is longer due to additional non-relevant terms, the relevance score of the longer document should not be higher than the shorter one. In a more recent study, LNC1 is analyzed in relation to its agreement with key retrieval axioms, particularly those addressing term frequency saturation and document length normalization [15].

mMARCO. The mMARCO dataset [3] plays a crucial role in addressing the gap of multilingual open-source information by overcoming the bias toward English-centric models and enabling a more robust evaluation of retrieval across diverse languages. As a multilingual benchmark dataset, mMARCO allows models such as ListConRanker [11] to leverage its extensive and diverse passages, facilitating global contrastive information learning and interaction within the ListTransformer.

3 Methodology

The methodology consists of 3 parts: creating the diagnostic dataset, running the activation patching experiments, and analyzing the results as internal mechanisms of the model [4].

3.1 Diagnostic Dataset

Chen et al. [4] study the TFC1 [7] axiom to analyze the term frequency information of query terms present in the documents. To perform activation patching, a diagnostic dataset is designed by creating (q, d_b, d_p) triples from the MS-MARCO developmental dataset [13] where one of the documents is in its original form, and the other is perturbed based on a randomly sampled query term. Next, the relevance scores from the model outputs are compared between the perturbed and the patched input runs to measure the impact of specific components [4]. The dataset is created based on the TFC1 axiom which is defined by Chen et al. [4] as:

TFC1. Let $q = w$ be a query with only one term w. Assume the length of document d_1 equals the length of document d_2. If the number of occurrences of w in d_1 is greater than the number of

occurrences of w in d_2, then for query q, the relevance score of d_1 should be higher than d_2.

Chen et al. [4] define two perturbations to observe the effects of TFC1: injection and replacement. TFC1-Inject (TFC1-I) and TFC1-Replace (TFC1-R) are outlined below:

TFC1-I. We randomly sample a query term and insert it at the end of the document d to create our perturbed document d_p. To create a baseline document d_b equal in length to our perturbed document, we insert a filler token (e.g., 'a') at the end of document d.

TFC1-R. We randomly sample one query term and replace all its occurrences in document d with a filler token to create a perturbed document d_p. The original document d acts as the baseline document d_b.

The dataset creation process can be seen in Figure 1. For TFC1-I, a randomly sampled query term is added either at the start, called prepend, or at the end of the document, called append. This document will now be the perturbed input. The other document in the triple is the baseline document, i.e. the original document, with filler tokens 'a' injected in the place of the added query term in the perturbed document. For TFC1-R, every appearance of the sampled query term is replaced by the filler token and the baseline document is just the original document. For each query, the 100 most relevant documents are retrieved and perturbed. For the final dataset, the 100 queries with the biggest average change of relevance score after perturbation were identified [4]. For our reproducibility experiments, we used the final version of the dataset given to us instead of reproducing the dataset due to the lack of details in the paper.

3.2 Activation Patching in Retrieval

The (q, d_b, d_p) triples from the diagnostic dataset are used to apply activation patching for the neural retrieval model. As seen in Figure 2, the patched run has X_{baseline} as input with cached activations coming from running the model with $X_{\text{perturbed}}$ as perturbations should lead to increases in relevance. This is the opposite for TFC1-R as that perturbation leads to a decrease in relevance Chen et al. [4]. Since we are working with a retrieval model, the dot product of query and document embeddings is used as a metric. The formula for measuring the difference between scores before and after patching is defined by Chen et al. [4] as:

$$\text{Patching Impact} = \frac{\text{Score}_{X_{\text{patched}}} - \text{Score}_{X_{\text{baseline}}}}{\text{Score}_{X_{\text{perturbed}}} - \text{Score}_{X_{\text{baseline}}}} \quad (1)$$

This formula measures how much (X_{patched}) shifts away from (X_{baseline}) and towards the relevance score of $(X_{\text{perturbed}})$. For TFC1-I, we are expecting the relevance score to increase after the perturbation. Therefore, the higher the metric value, the more important the activation is to the result. Based on the axiom, the patching should show how and where the term frequency information of the selected query term is stored.

4 Extension

We propose two extensions to validate the results from the original paper and further analyze the TAS-B model's behavior. First, we perform causal interventions based on the LNC1 axiom to understand
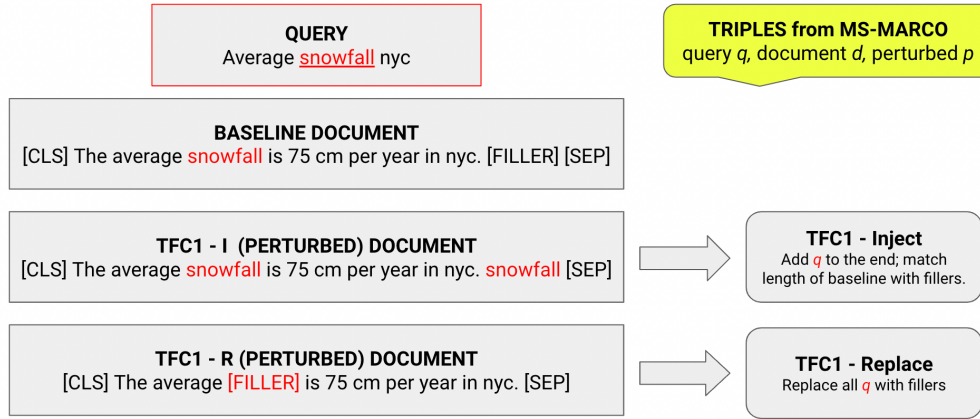
Interpreting Multilingual and Document-Length Sensitive Relevance Computations in Neural Retrieval Models through
Axiomatic Causal Interventions

SIGIR '25, July 13–18, 2025, Padua, Italy

Figure 1: Diagnostic dataset creation setup for TFC1-I and TFC1-R.
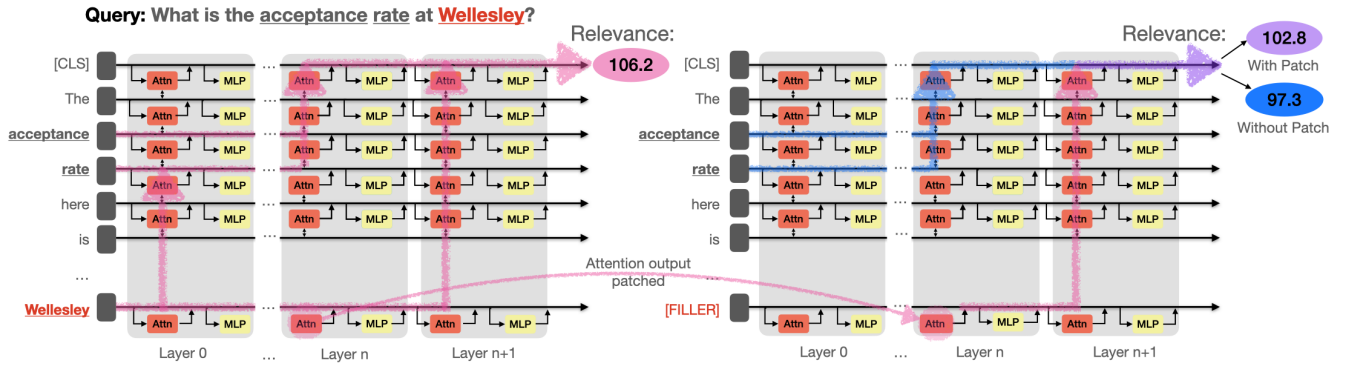

Figure 2: Activation patching setup for retrieval from the original paper. In this example, a (q, d_b, d_p) triple is created to measure term frequency effects in the model. A perturbed document (left) is created by injecting the sampled query term ("Wellesley") at the end, and a baseline document (right) is made by adding filler tokens. When running the model with the perturbed input, the activations are saved. For the patched run, the model runs on the baseline document, and the cached activations are patched in. This should lead to an increase in the ranking score. [4]

how the model encodes document-length information. Secondly, we use the mMARCO Spanish and Chinese language datasets to understand how the term frequency localisation compares between different languages.

4.1 LNC1

To better understand the internal mechanisms of the retrieval model, we apply axiomatic causal interventions based on another axiom, namely LNC1. The axiom can be described as the following [7]:

LNC1. Let q be a query and d_1, d_2 be two documents. Assume that for some word $w' \notin q$, $c(w', d_2) = c(w', d_1) + 1$. If for any query term w, $c(w, d_2) = c(w, d_1)$, then for query q, the relevance score of d_1 should be greater than or equal to d_2.

This principle can be used to evaluate how effectively neural retrieval models handle document-length variability, a critical aspect in datasets like Touché 2020 [2], which often contain noisy, short, and non-argumentative documents. While traditional models such as BM25 show near-perfect adherence to LNC1, neural

models like TAS-B [10] and SPLADEv2 [9] frequently violate this axiom, leading to a suboptimal ranking of documents with varying lengths. This analysis reveals that LNC1 serves as a diagnostic tool to pinpoint specific weaknesses in neural retrieval systems, such as their inability to prioritize longer, more detailed arguments over shorter, less informative ones.

We are using a similar experimental setup as the original paper, with a few modifications to analyze the length property. The variants of prepending and appending query terms are not relevant for the LNC1 experiments. Instead, we add terms that are not relevant to the query to decrease relevancy. At first, we experimented with adding a single filler token. However, that did not lead to a consistent enough decrease in the relevance score. The same behaviour has been observed with the use of a token not present in the queries, such as "guantanamo". This eventually leads us to doubling the document size with the injected tokens.

Moreover, both inputs must have the same length. To achieve this, we tested masking the injected tokens, which resulted in a
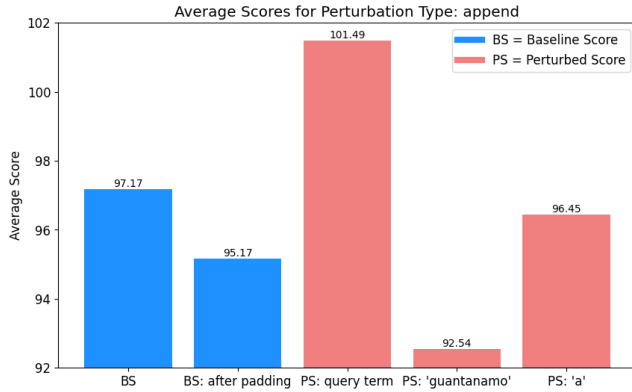
Figure 3: The average relevance score comparison shows the baseline score with and without padding. The perturbed conditions include appending a query term, 'guantanamo' tokens, and 'a' tokens.

clear decrease in relevance. As seen in Figure 3, the most consistent way of achieving a decrease in relevance after the perturbation is to use padding tokens for X_{baseline} and 'guantanamo' tokens for $X_{\text{perturbed}}$, as 'guantanamo' doesn't appear a single time in the documents and queries in the final dataset.

The experiment involves inserting 'guantanamo' tokens to construct $X_{\text{perturbed}}$, and padding tokens in place of them for X_{baseline}. Next, activations from the perturbed run are saved and used to patch activations when running the model with the padded baseline document. This approach differs from the original paper, as their saved activations are typically from the higher-relevance run. However, given the design of the ranking metric used, the patched activations should still highlight differences in length information.

4.2 mMARCO Dataset - Spanish and Chinese

In this extension, we want to determine whether the conclusions drawn from the previous experiment on an English-language dataset hold across different languages. To achieve this, we use the mMARCO dataset [3], the multilingual version of MS MARCO, which includes 13 languages, excluding the original English version. We experiment with Spanish and Chinese languages to see if fundamental differences between languages impact the behaviour of the model.

The experimental setup is similar to MS MARCO in the original paper. To find the top 100 relevant documents, we only use documents from the corpus associated with queries instead of going through the entire corpus. For the final dataset, we chose the 100 queries with the highest relative change from the append perturbation as Chen et al. [4] did not specify which perturbation was used to select the queries.

5 Experimental Setup

The model used in the original work is TAS-B [10], a DistilBERT-based model with 6 layers and 12 attention heads per attention layer. TAS-B is a bi-encoder, which processes queries and documents separately and uses a pooled representation of the CLS token for ranking score calculation. The choice of the model was due to

its high performance in ranking tasks despite its simplified architecture. [4]

Activation patching is done in multiple parts of the transformer: (1) residual stream, (2) attention outputs, (3) MLP outputs, and (4) individual attention heads. The first 3 are done by applying patching for each token in the document, which is referred to as the block experiment. Most of the results are based on the TFC1-I appending perturbation [4].

The patching experiments quantify the impact of each component in terms of term frequency information. The tokens are classified as per Table 1. The darker the shade of blue in the square of the heatmap, the more significant the impact of the activations at those tokens is in increasing performance.

Table 1: Token type classifications for documents. TFC1-I perturbed documents include all six token types, while TFC-R perturbed documents have five token types since no terms are injected during perturbation.

Label	Definition
tok_{CLS}	The CLS token.
tok_{inj}	The selected query term injected into the document.
$\text{tok}_{\text{qterm}^+}$	Occurrences of the selected query term that already exist in the original document.
$\text{tok}_{\text{qterm}^-}$	Occurrences of the non-selected query terms in the original document.
$\text{tok}_{\text{other}}$	Terms in the original document that are not query terms.
tok_{SEP}	The SEP token.

6 Experimental Results

This section describes the results of reproducing the original paper and the two extensions we designed. Similar to the experiments in [4], we report results from a single run. A fixed random seed is set, and we use a batch size of 1.
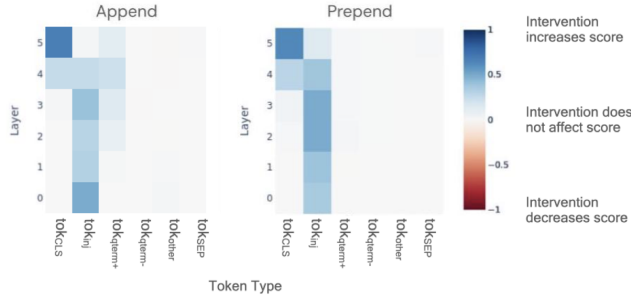
6.1 Reproduced Results

6.1.1 Importance of Added/Deleted Query Terms (The Block Experiment). This experiment investigates how specific activations within the residual stream and attention outputs can help identify the tokens that have the most significant impact on performance.

Figure 5 shows the reproduced results of the append block experiment, while Figure 6 shows the reproduced results of the prepend block experiment. For comparison, Figure 4 presents the original results of the block experiment with append and prepend on the residual stream.

Append Both the original and reproduced heatmaps exhibit strong activations for the injected query term and the pre-existing occurrences of the selected query term in Layer 4, indicating that the existing selected query term occurrences remain relevant in determining the overall performance. Following this, the term frequency information is passed to the CLS token, which shows the strongest activation in Layer 5.

However, slight variations are observed between the two heatmaps. In Layer 4, the original append heatmap demonstrates a similar

Interpreting Multilingual and Document-Length Sensitive Relevance Computations in Neural Retrieval Models through
Axiomatic Causal Interventions

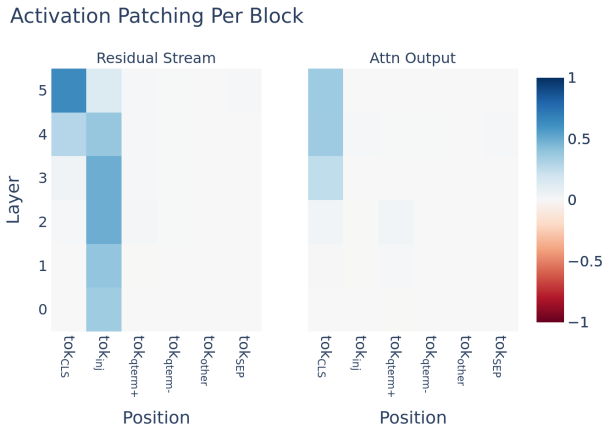SIGIR '25, July 13–18, 2025, Padua, Italy

Figure 4: The original block experiment with append and prepend from patching into the residual stream.
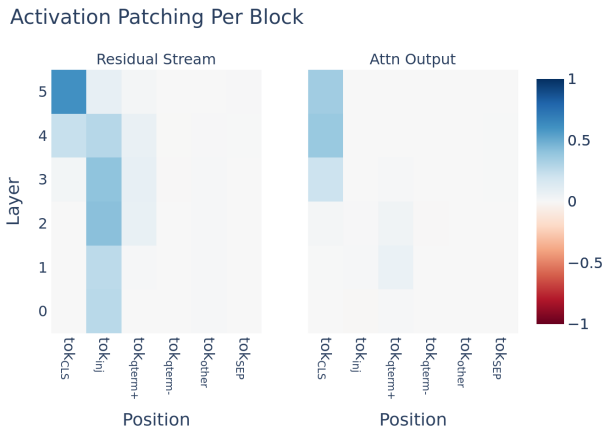
Figure 5: The reproduced block experiment results for append from patching into the residual stream and attention output.

Figure 6: The reproduced block experiment results for prepend from patching into the residual stream and attention output.

impact on the CLS and injected tokens, whereas the reproduced heatmap indicates a stronger impact for the injected tokens compared to the CLS token on the overall performance. Additionally, in Layer 5 of the original heatmap, alongside the strongest impact of the CLS token, a weak impact is visible from the pre-existing occurrences of the selected query term. In contrast, the reproduced heatmap shows this weak impact coming from the injected tokens. Overall, there are slight differences in the impact of some tokens between the original and reproduced heatmaps but they reflect a similar picture.

Prepend It was suggested by Chen et al. [4] that the model stores term frequency information in query term positions that are present at the beginning of the document. To verify this, the prepend experiment was done. For this experiment, the heatmaps exhibit a strong impact for the injected query term tokens and show that by Layer 4 the term frequency information is shifted to the CLS token with the strongest impact in Layer 5. However, the original claim about term frequency information being stored at the beginning of the document is not as strongly supported by the updated results, as the strongest activations appear for the injected tokens as opposed to the first appearance of the query term. However, the impact of the injected tokens for the prepend experiment is indeed stronger than for append, and only the append heatmap shows the impact of non-injected query terms as they are closer to the beginning of the document.

Additionally, we have included the heatmaps of the attention outputs for append and prepend that were missing from the original report. For attention outputs, there is a lower impact at later layers at the CLS token position compared to the residual stream. We did not include MLP output results as they do not contain any significant information.

6.1.2 Activation Patching on Individual Attention Heads. Based on the previous results, Chen et al. [4] hypothesize that specific attention heads are primarily responsible for processing and carrying the term frequency information throughout the model. To test this hypothesis, activation patching was conducted on the individual attention heads. Figure 7 presents the original results of the attention head experiment with append, showing the results of the top and bottom 10% relevant documents per query for append. The results with prepend were not mentioned or included in the paper. The results of TFC1-R are mentioned as not relevant as the attention heads amplify an existing relevance signal.

Append Figure 8 shows the reproduced results of activation patching on individual attention heads experiment with append for the top 10% of relevant documents per query. Both the original and reproduced results for the top 10% might differ slightly in numbers but reveal the same 4 attention heads that encode the TFC1 axiom, i.e. (Layer 0. Head 9), (1.6), (2.3), (3.8). As no significant findings were reported based on the bottom 10 % of relevant documents, they have been excluded in our paper.

Prepend The results for patching attention head activations with prepend were not reported in the original paper. Based on the results produced by us in Figure 9, this was because the values across heads are closer to zero for the top-ranked documents, contrary to what we see for append. Similarly, values are close to zero for the bottom-ranked documents as well. Interestingly, the 3 attention
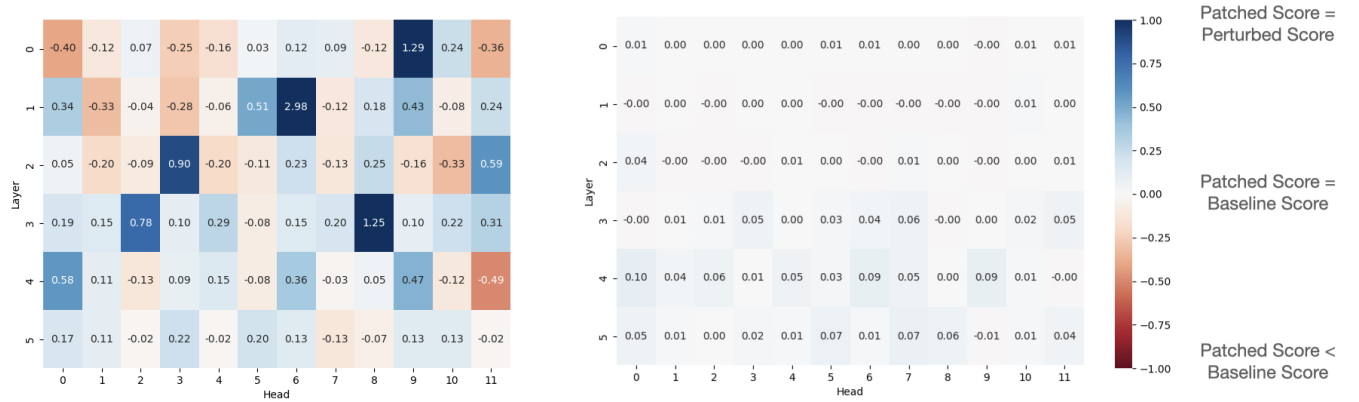
Figure 7: The original results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top(left) and bottom(right) 10% relevant documents per query for append.
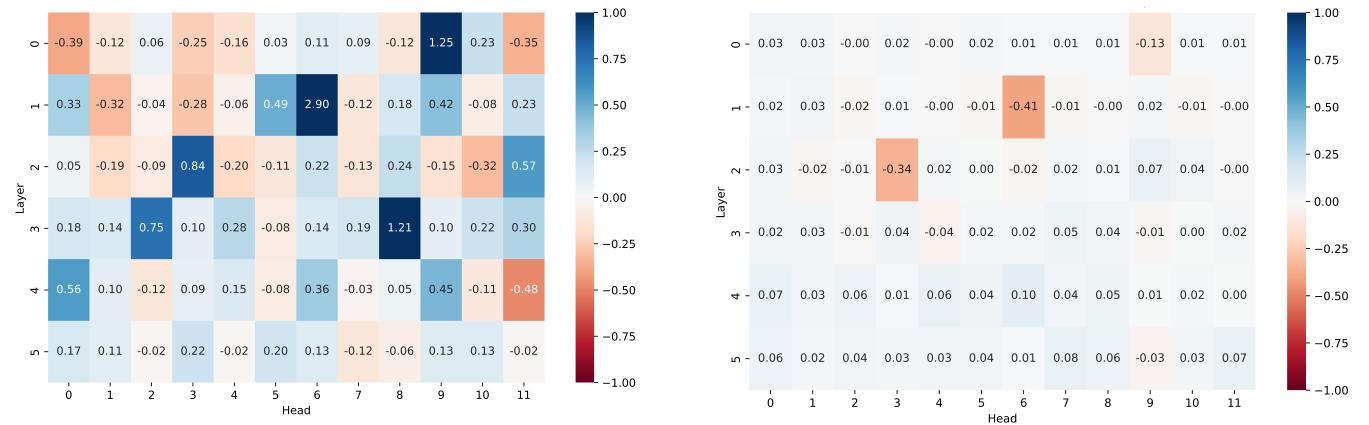

Figure 8: The reproduced results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for append.

Figure 9: The reproduced results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for prepend.

heads with the lowest impact were from the list of 4 attention heads with the strongest impact that were identified for the appending perturbation. Therefore, including these results in the appendix should have been done to give more depth to the findings.

6.2 Extension Results

6.2.1 ***LNC1***. Figure 10 shows the results of the append block experiment in a heatmap that is similar to the original paper. We report one less column, as in our experiment, we are not injecting query terms. Although this time the patching activation method should decrease relevance, due to the ranking equation, darker blue still represents the most significant impact.

The append block experiment shows no impact on any token other than the CLS token in the subsequent layers. This is to be expected as document length is not directly connected to any type of token, including the injected ones, as BERT stores sequence-level information in the pooled representation of the CLS token.

Furthermore, Figure 11 supports these findings, as injecting tokens at the beginning seems to lead to similar results.

Figure 12 shows the attention head results of the top 10% ranked documents for append. The attention head patching has different results from the original paper because, for append, almost all attention heads have a mildly positive impact.

Figure 13 shows the top 10% ranked documents for prepend. There are 2 heads that stand out from the rest for their positive impact and 2 that seem to stand out for their negative impact. None of these values are comparable to the values of attention heads identified in the original paper.

6.2.2 ***mMARCO - Spanish and Chinese***. The results of the mMARCO Spanish and Chinese block experiments, shown in Figures 14 and 15, closely resemble those of the original English experiment. However, the values for injected and the CLS token appear to have a slightly lower impact on ranking for Spanish, while other occurrences of the query term show minimal importance. The prepend
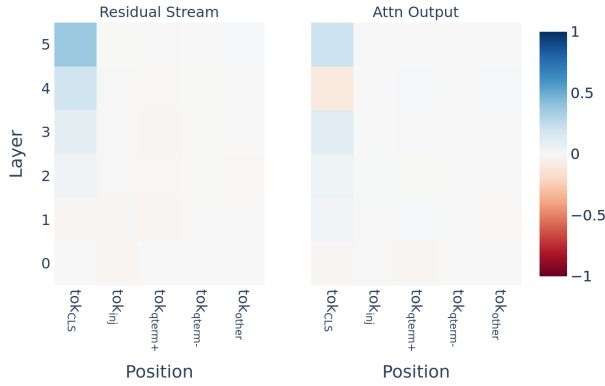
Interpreting Multilingual and Document-Length Sensitive Relevance Computations in Neural Retrieval Models through Axiomatic Causal Interventions
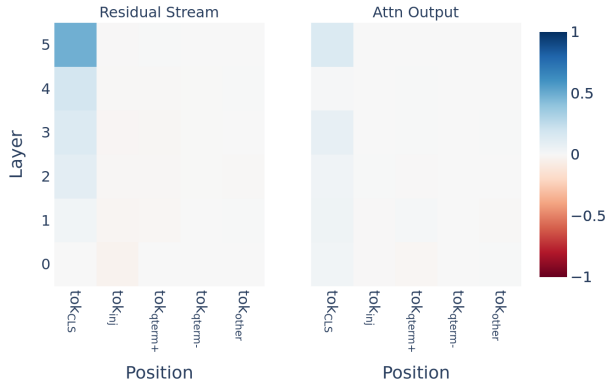
SIGIR '25, July 13–18, 2025, Padua, Italy

Figure 10: LNC1 block experiment results for append from patching into the residual stream and attention outputs.

Figure 12: LNC1 results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for append.

Figure 11: LNC1 block experiment results for prepend from patching into the residual stream and attention outputs.

Figure 13: LNC1 results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for prepend.

results in Figure 16 and 17 look similar to the original experiments. Therefore, this extension supports the claim that term frequency is stored across occurrences of the query term, across languages as well. The impact of the injected tokens is slightly higher for prepend, which also shows some support for the information mostly being localized near the beginning of the document.

Figures 18 and 19 show the results of patching attention heads for the top 10% ranked documents. Compared to the original paper, these results are different because for both prepend and append, the impact of each head seems to be very small and no high-impact heads can be identified. This might be because the model was pre-trained on the English language so there aren't clear attention patterns developed by this model for other languages.
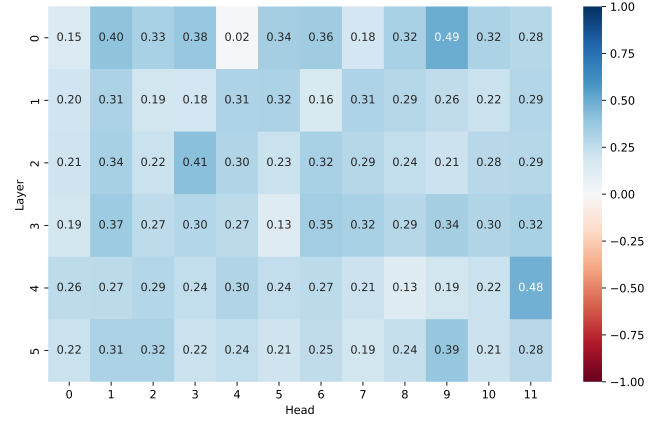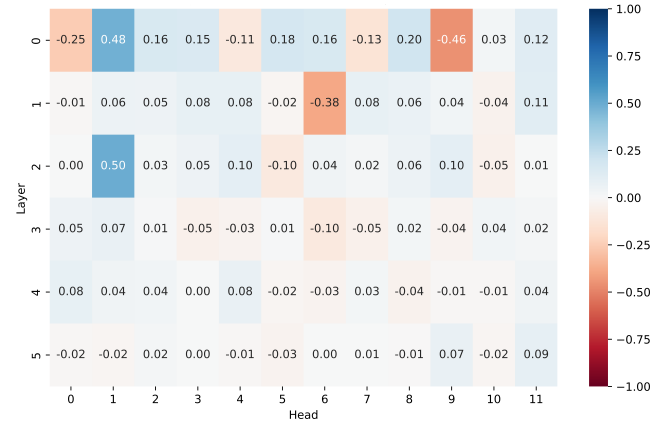
7 Discussion

The original paper "Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models" [4] investigated the internal mechanisms of a transformer-based neural retrieval model. It introduces a novel dataset and activation patching variation suitable for information retrieval. The results of their analysis show that query term frequency information is stored in specific tokens, while at different layers, a few key attention heads play a crucial role in encoding this information. In certain experiments, particularly those involving TFC1-I appending, these tokens and heads are strongly identifiable, demonstrating the effectiveness of the applied method.
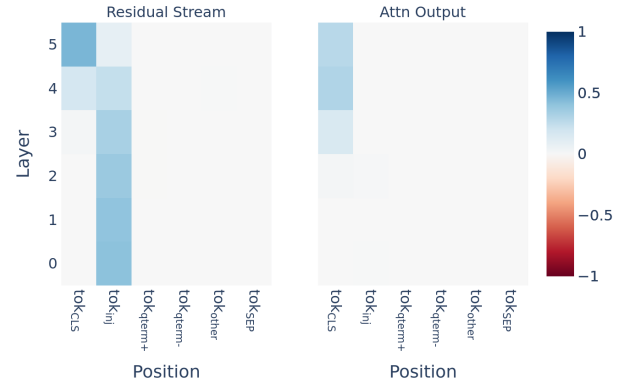
Our extension testing of the LNC1 axiom showed that, similar to the TFC1 axiom, the information about document length is
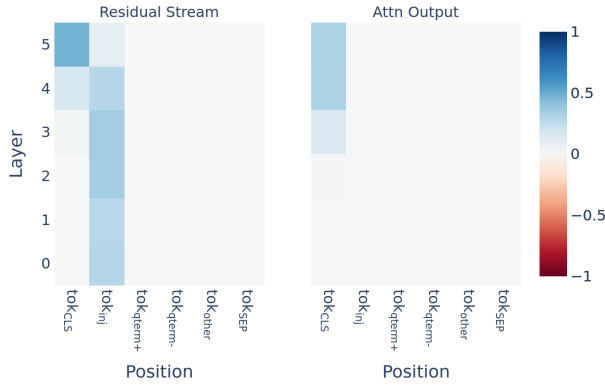
Figure 14: TFC1 mMARCO Spanish language block experiment results for append from patching into the residual stream and attention outputs.
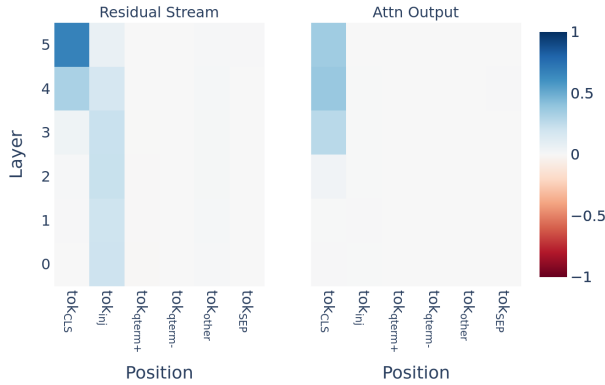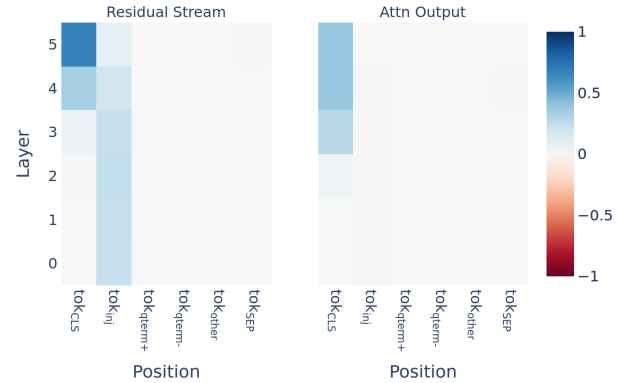

Figure 16: TFC1 mMARCO Spanish language block experiment results for prepend from patching into the residual stream and attention outputs.

Figure 15: TFC1 mMARCO Chinese language block experiment results for append from patching into the residual stream and attention output.
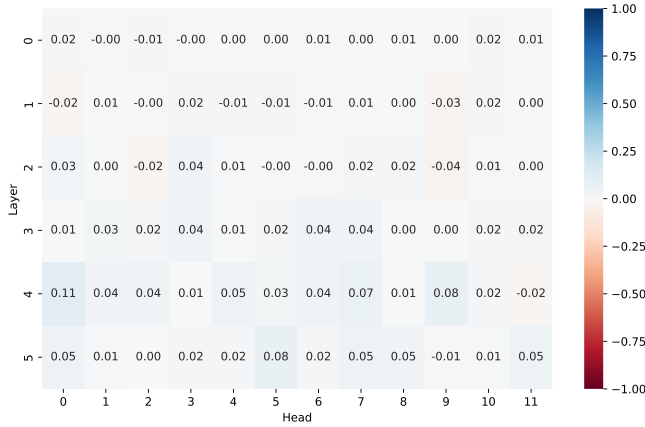

Figure 17: TFC1 mMARCO Chinese language block experiment results for prepend from patching into the residual stream and attention output.

stored at the CLS token in the later layers of the model. Additionally, experiments using Spanish and Chinese datasets confirmed that term frequency information is stored across term occurrences, though the evidence for a stronger signal near the beginning of the document was only marginal. At the same time, neither extension leads to identifying specific attention heads, which is especially unexpected for the Spanish and Chinese language data where the perturbations made are exactly the same kind. This could mean that attention in this model works differently for other languages compared to English. While [4] included an ablation test to further prove the claim, the fact that these results were only achieved by appending a query term for the top-ranked documents, and not for any other type of perturbation, means that the value of these attention heads should be studied more to get a better understanding of the model.

During the creation of datasets for the extension experiments, we observed the importance of carefully selecting queries and documents where the model correctly increases or decreases the relevance score after axiom-based perturbations. This is due to the fact that the initial activation patching setup requires the model to almost always follow the rules coming from the axiom. However, this is not the case with both the TFC1 and LNC1 axioms.

Additionally, even if we assume the model takes these properties into account, unexpected cases are not accounted for in the design of the ranking metric. It always assumes that the perturbation has the same kind of impact, which in the reported results of the original paper, should always be increasing the relevance score. However, in cases where the baseline score is higher than the relevance score after perturbation, the patching score is most likely

Interpreting Multilingual and Document-Length Sensitive Relevance Computations in Neural Retrieval Models through Axiomatic Causal Interventions

SIGIR '25, July 13–18, 2025, Padua, Italy

Figure 18: TFC1 mMARCO Spanish language results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for append.
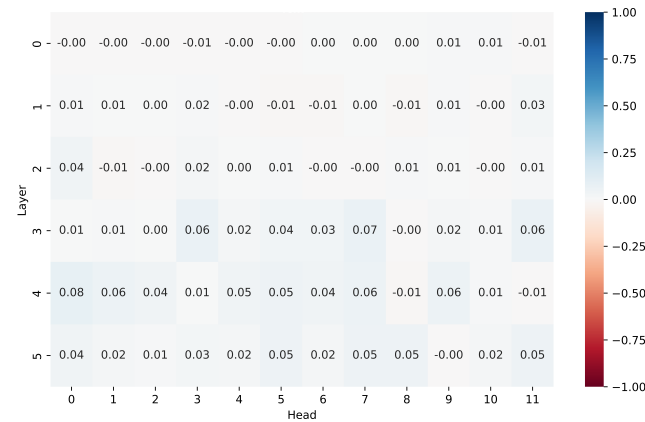
Figure 19: TFC1 mMARCO Chinese language results of activation patching on individual attention heads experiment to find the attention heads that encode the TFC1 axiom showing results of the top 10% relevant documents per query for append.

to fall between the baseline and perturbed score if the patched module has an impact. This issue became apparent when we initially ran extension experiments on datasets that did not consistently produce axiom-compliant behaviour. As a result, some findings contained a mix of positive and negative scores, making interpretation difficult. Addressing these cases requires further investigation, and potentially, a more adaptive ranking metric should be developed to capture variations more accurately.

8 Conclusion

In this paper we reproduce and extend "Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models", aiming to validate their original findings. We

successfully reproduced most of the key experiments, revealing that the TAS-B model indeed stores term frequency information across occurrences of the query term, possibly more at the start of the document. Additionally, we could see that the neural retrieval model moves this information to the CLS token pooled representation in the later layers. Our LNC1-based experiments confirmed the importance of the CLS token for encoding document-level features while our extension with the Spanish and Chinese mMARCO datasets showed that some of these term-frequency mechanisms generalize beyond English.

In summary, this work reaffirms that the designed activation patching method can uncover specific mechanisms of relevance modelling, such as term frequency and document length encoding. However, it also shows the need for more robust interpretability approaches and more focus to guarantee the reproducibility of machine learning research. Future research could focus on investigating circuits of components instead of singular modules to obtain a clearer picture of how information is processed while also further looking into some of the findings of this paper. This could include testing different ranking metrics and perturbations to gain a wider knowledge of the inner workings of neural retrieval models. Perhaps, that will assist in identifying specific mechanisms across other tasks and languages.

References

[1] Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. arXiv:2102.12452 [cs.CL] https://arxiv.org/abs/2102.12452

[2] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval: Extended Abstract. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings* (Thessaloniki, Greece). Springer-Verlag, Berlin, Heidelberg, 384–395. https://doi.org/10.1007/978-3-030-58219-7_26

[3] Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. arXiv:2108.13897 [cs.CL] https://arxiv.org/abs/2108.13897

[4] Catherine Chen, Jack Merullo, and Carsten Eickhoff. 2024. Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models. arXiv:2405.02503 [cs.IR] https://arxiv.org/abs/2405.02503

[5] Tanya Chowdhury and James Allan. 2024. Probing Ranking LLMs: Mechanistic Interpretability in Information Retrieval. arXiv:2410.18527 [cs.IR] https://arxiv.org/abs/2410.18527

[6] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600 [cs.LG] https://arxiv.org/abs/2309.08600

[7] Hui Fang, Tao Tao, and ChengXiang Zhai. 2004. A formal study of information retrieval heuristics. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, United Kingdom) *(SIGIR '04)*. Association for Computing Machinery, New York, NY, USA, 49–56. https://doi.org/10.1145/1008992.1009004

[8] Hui Fang, Tao Tao, and ChengXiang Zhai. 2011. Diagnostic Evaluation of Information Retrieval Models. *ACM Trans. Inf. Syst.* 29 (04 2011), 7. https://doi.org/10.1145/1961209.1961210

[9] Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. arXiv:2109.10086 [cs.IR] https://arxiv.org/abs/2109.10086

[10] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. arXiv:2104.06967 [cs.IR] https://arxiv.org/abs/2104.06967

[11] Junlong Liu, Yue Ma, Ruihui Zhao, Junhao Zheng, Qianli Ma, and Yangyang Kang. 2025. ListConRanker: A Contrastive Text Reranker with Listwise Encoding. arXiv:2501.07111 [cs.CL] https://arxiv.org/abs/2501.07111

[12] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *arXiv preprint arXiv:2202.05262* (2022). https://arxiv.org/abs/2202.05262

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated Machine Reading Comprehension Dataset. https://microsoft.github.io/msmarco/ Choice 2640 (2016), 660.

[14] Andrew Parry, Catherine Chen, Carsten Eickhoff, and Sean MacAvaney. 2025. MechIR: A Mechanistic Interpretability Framework for Information Retrieval. arXiv:2501.10165 [cs.IR] https://arxiv.org/abs/2501.10165

[15] Nandan Thakur, Luiz Bonifacio, Maik Fröbe, Alexander Bondarenko, Ehsan Kamalloo, Martin Potthast, Matthias Hagen, and Jimmy Lin. 2024. Systematic

Evaluation of Neural Retrieval Models on the Touché 2020 Argument Retrieval Subset of BEIR. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 1420–1430. https://doi.org/10.1145/3626772.3657861

[16] Fred Zhang and Neel Nanda. 2024. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. arXiv:2309.16042 [cs.LG] https://arxiv.org/abs/2309.16042