# Can Embeddings Analysis Explain Large Language Model Ranking?


Claudio Lucchese
claudio.lucchese@unive.it
Ca' Foscari University of Venice
Italy

Giorgia Minello
giorgia.minello@unive.it
Ca' Foscari University of Venice
Italy

Franco Maria Nardini
francomaria.nardini@isti.cnr.it
ISTI-CNR
Italy

Salvatore Orlando
orlando@unive.it
Ca' Foscari University of Venice
Italy

Raffaele Perego
raffaele.perego@isti.cnr.it
ISTI-CNR
Italy

Alberto Veneri
alberto.veneri@unive.it
Ca' Foscari University of Venice
ISTI-CNR
Italy



## ABSTRACT

Understanding the behavior of deep neural networks for Information Retrieval (IR) is crucial to improve trust in these effective models. Current popular approaches to diagnose the predictions made by deep neural networks are mainly based on: i) the adherence of the retrieval model to some axiomatic property of the IR system, ii) the generation of free-text explanations, or iii) feature importance attributions. In this work, we propose a novel approach that analyzes the changes of document and query embeddings in the latent space and that might explain the inner workings of IR large pre-trained language models. In particular, we focus on predicting query/document relevance, and we characterize the predictions by analyzing the topological arrangement of the embeddings in their latent space and their evolution while passing through the layers of the network. We show that there exists a link between the embedding adjustment and the predicted score, based on how tokens cluster in the embedding space. This novel approach, grounded in the query and document tokens interplay over the latent space, provides a new perspective on neural ranker explanation and a promising strategy for improving the efficiency of the models and Query Performance Prediction (QPP).



## CCS CONCEPTS

• **Information systems** → *Language models*; • **Computing methodologies** → **Ranking**.

## KEYWORDS

Explainable Artificial Intelligence, Large Language Models, Embeddings Analysis, Text Ranking

**ACM Reference Format:**
Claudio Lucchese, Giorgia Minello, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Alberto Veneri. 2023. Can Embeddings Analysis Explain Large Language Model Ranking?. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615225





*CIKM '23, October 21–25, 2023, Birmingham, United Kingdom*



ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615225


## 1 INTRODUCTION

Simple yet effective Neural Information Retrieval (NeuIR) models derived from Large Language Models (LLMs), are increasingly used in state-of-the-art retrieval strategies. As an example, the leaderboard of the well-known MS MARCO [10] passage ranking contest is currently dominated by methods that employ re-ranking models based on LLMs, such as BERT [12], RoBERTa [8], and ELECTRA [1]. In this work, we are interested in analyzing the prediction phase of cross-encoder models, more specifically, the class of models also known as Mono-X models [22]. In particular, we focus on the role of the embeddings and their spatial placement in the latent space to understand how a cross-encoder model behaves during the ranking score prediction phase and to unveil if this information can be used to estimate the quality of the prediction without knowing the ground truth, i.e., for QPP.

One of the common and most used methods to explain predictions of Machine Learning (ML) models for IR is to assign an *importance score* to each feature of an input instance. In this category of explanations, we can cite methods like EXS [17], LIRME [20], and DeepSHAP for NeuIR model [4]. We should consider this approach a "shallow" explanation method because it only provides a score assessing the importance of a certain token, leaving unresolved the question of "why" a model has arrived to give more importance to a document rather than to another one. The explanation via *query intent modeling* is another approach specific to the IR field. The intent modeling can be done via query expansion [18] or by generating a verbose query description [24]. However, this type of explanation is more beneficial for the end-user than the ML developers since it is difficult to understand if the explanation is faithful to the model behavior. Another type of explanation for the IR field is the explanation by *proving IR axioms*. Examples of such research direction are the work of Rennings et al. [15] and of Câmara and Hauff [2]. Even though these works aim to answer the same question that we address, we claim that such axioms are reasonable for humans to explain the relevance of a ranked query/document pair, but in general, what applies to humans does not directly apply to machines, as discussed in [2]. Like various other works, we instead present explanations through *model analysis*. In this approach, the

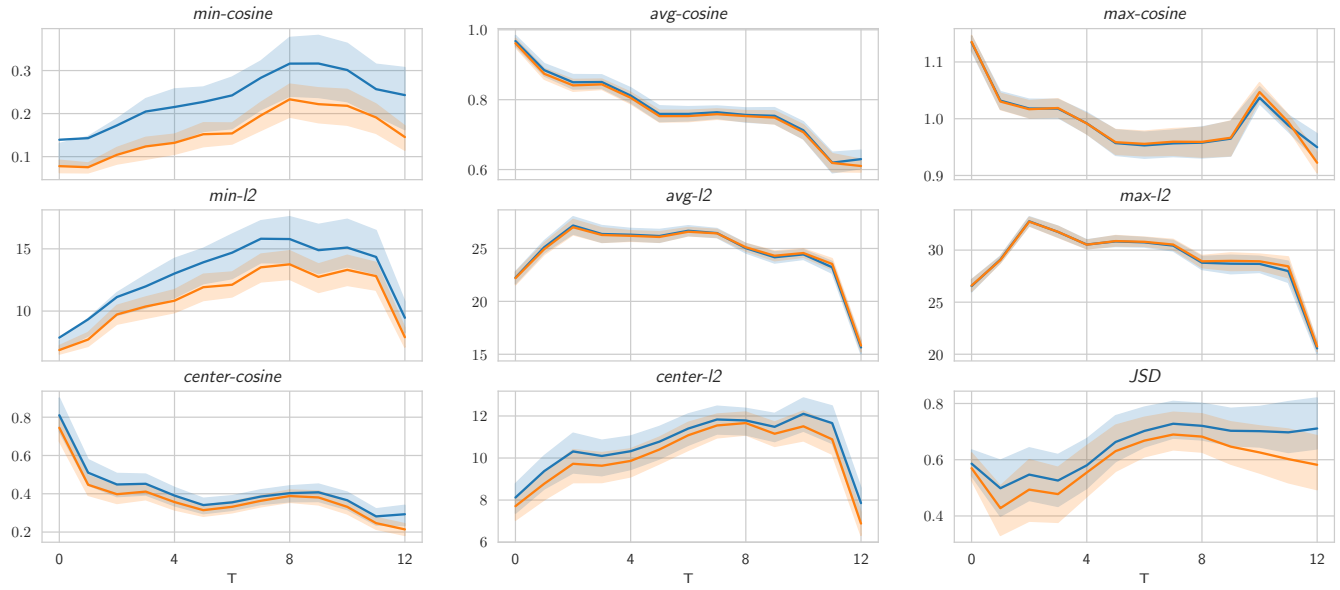



**Figure 1: Distance between the mean of query/document pairs with a high score (in orange) and query/document pairs with a low score (in blue) using all the metrics with *cosine*, *l2*, and *JSD*. The shadowed area represents the interquartile range.**

NeuIR model is analyzed to understand if some of its characteristics can be associated with a specific human cognitive way of reasoning or to describe the inference strategy. This kind of analysis is mostly conducted on the attention layers to understand their role in the final prediction. Examples of such approaches in the IR field are presented in [23] and in [13]. Since we consider the geometry of the embedding space rather than the attention layers, the closest works to our are [3] and [14], where the authors analyze the geometry of LLMs, and find interesting insights about the contextualization of words. However, their works do not consider, as we do, some peculiar aspects related to the ranking problem, such as the relationship between query tokens and document tokens and between score and relevance judgments.

## 2 EMBEDDING ANALYSIS IN MONO-X MODELS

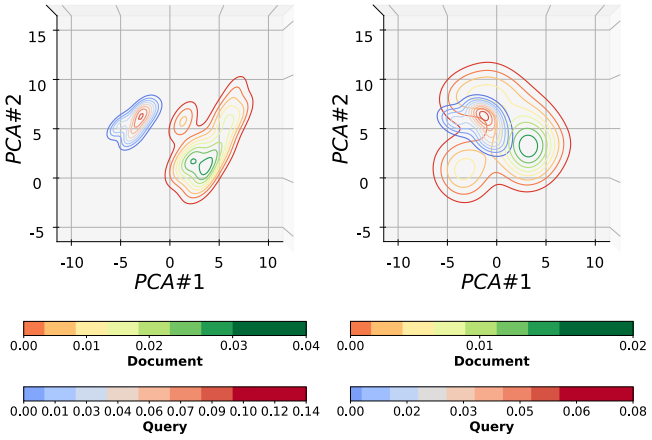
Our work is based on an intuition derived from analyzing the main components of LLMs based on transformer blocks, i.e., the attention mechanism [19]. It might seem intuitive to explain an LLM-ranker's prediction using the information stored in the attention layers. However, the role of attention in explaining the prediction of a LLM-based NeuIR model is currently not well-defined. There are works claiming that attention is not really useful to be used as an explanation [5, 16], while others claim that it might be helpful in some scenarios [21]. In this work, we analyze the inference phase of such models from a different perspective. The attention mechanism commonly used is self-attention, which can be seen as a simple weighted sum of the embeddings in input in the attention layer since the *queries* and the *keys* used are the same [11]. We recall that inside a LLM such as BERT we have multiple self-attention heads for each transformer block. If we have a uniform distribution of the attention weights in every attention head of each transformer block and we assume that other parts of the network (e.g., the fully-connected layers) do not change the embeddings too much, we should observe tokens passing through the various blocks and asymptotically converge to a single point.

**Research questions.** Given the aforementioned intuitions and motivations, we formulate the following research questions:

- RQ1. Do embeddings of a LLM-based neural ranker follow a common pattern in their movement in the latent space during the traversal of the various transformers block?
- RQ2. Can we use the geometry information from the embeddings to understand whether the predictions are accurate?

Unlike previous approaches, by RQ1, we focus on the embedding movements in the latent space and investigate if we can recognize meaningful patterns that can be useful for model debugging and/or model efficiency improvement. Instead, RQ2 focuses on finding a correlation between the score produced, the true relevance of the query/document pair, and the arrangement of the embeddings in the latent space. If present, it can be very interesting for various IR open problems, including, for instance, QPP.

**Dataset and models used.** Our analysis employs different versions of the Microsoft Machine Reading Comprehension Dataset (MS MARCO) for passage re-ranking. Specifically, we use the versions provided by *ir-dataset* [9]. We refer to specific dataset versions by using their id in *ir-dataset*. To fine-tune our models, we used the version containing only queries with at least one relevance judgment (*msmarco-passage/train/judged*). To analyze our models, we used the query relevance judgment from the *msmarco-passage/trec-dl-2019/judged* and *msmarco-passage/trec-dl-2020/judged* versions of the dataset. Since we are interested in comparing the score of the model w.r.t. the real relevance judgments, we considered for each query only the documents having an associated relevance judgment.




(a) Low Score (b) High Score

**Figure 2: PDF of query and document tokens (via PCA) represented with contour lines for both low-score (a) and high-score (b) instances at the last transformer block ($T_{12}$). Ref. *trec-dl-2019* - Query ID: 130510, Document ID (low score): 2061706, Document ID (high score): 8612903.**

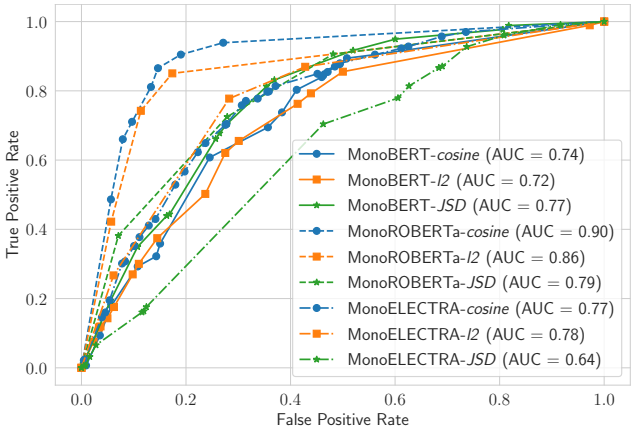



**Figure 3: Area Under the Curve (AUC) over the test set for each model and metric under consideration. The false positive rate represents the rate of query/document pairs wrongly classified as misprediction.**

In this work, we analyze three popular NeuIR models based on three different LLMs considered among the state-of-the-art methods for text ranking: BERT, RoBERTa [8], ELECTRA [1]. We used the small version of each model with 12 transformer blocks. We fine-tuned[1] the pre-trained models following [12]. The only adjustment made during the fine-tuning is related to the batch size, which was reduced to fit our GPU memory. Specifically, we used batch sizes of 64 and 32. The performance of our models in terms of *MRR* with a cutoff at ten over the official small development set (*msmarco-passage/dev/small*) are 0.337 for MonoBERT, 0.340 for MonoRoBERTa and 0.347 for MonoELECTRA, aligned with the results available in the literature [7, 22].

## 3 ANALYZE THE DISTANCE BETWEEN QUERY AND TOKEN EMBEDDINGS

To answerRQ1 we apply a simple and intuitive analysis focused on the spatial relation of query and document tokens. To this end, we label query tokens and document tokens as forming in the embedding space two distinct classes of tokens $S_q$ and $S_d$. We highlight that, in almost all the experiments in this study, we discard all the special tokens, like, for example, the tokens *[SEP]*, *[CLS]*, and *[PAD]* in *MonoBERT*.

We base our analysis on three different distance measures between the two groups of tokens: cosine distance (*cosine*), Euclidean distance (*l2*), and Jensen–Shannon Divergence (*JSD*) distance. For *cosine* and *l2*, we compute the maximum, minimum, and average values between the tokens in $S_q$ and $S_d$ (identified with the prefixes *max*, *min*, and *avg* respectively), and we also consider the distance from the centroids of $S_q$ and $S_d$ (identified with the prefix *center*). Concurrently, we compute *JSD* distance by modeling the probability distribution over $S_q$ and $S_d$ as follows. First, we project the token high-dimensional space on a lower one by Principal Component Analysis (PCA). To perform this conversion, we used all the embeddings (query, documents, special tokens, and padding) as input data. Then, by the first two principal components, we compute the Probability Density Function (PDF) of both query and document token sets via Kernel Density Estimation (KDE), particularly suitable for our purpose being a non-parametric method. Finally, we measure the distance between these two distributions through the square root of the *JSD*. Let $\phi_d$ and $\phi_q$ be the two distributions, then the *JSD* is defined by $JSD(\phi_d||\phi_q) = H\left(\frac{\phi_d+\phi_q}{2}\right) - \frac{H(\phi_d)+H(\phi_q)}{2}$, where $H(\cdot)$ is the Shannon entropy.

We analyze the aforementioned measures w.r.t. each transformer block $T$, starting from $T_0$ until $T_{12}$. $T_0$ represents the initial input embeddings fed into the neural network while $T_{12}$ represents the last transformer block of each architecture under analysis (we use the small version of each LLMs under consideration). For the sake of brevity, we present the results obtained on the *trec-dl-2019* dataset and only for MonoBERT. The results comparing high and low score embedding trends for *cosine*, *l2*, and *JSD* distance are shown in Figure 1. It is clear from the plots that there is a difference in the behavior of tokens belonging to high-scored (greater than 0.5) and low-scored (less than 0.5) documents, especially for *min-cosine*, *min-l2* and *JSD*. To confirm the intuition from the plots, we perform a permutation test with the assumption that the data comes from the same distribution as null hypothesis and a one-side alternative hypothesis where we hypothesize that the mean of the high-scored documents is higher with respect to the mean of the low-scored documents. The results of the statistical test, setting a p-value limit to 0.01, show that the difference is statistically significant for *min-cosine*, *center-cosine*, *min-l2*, *center-l2* and *JSD* in all the transformer blocks. We highlight that a difference between the token distance is expected. According to the common intuition that the token is contextualized during the transformation in each transformer block, i.e., we expect that query tokens are contextualized in the document tokens when the score is high. However, quite unexpectedly, this

[1]Code available at: https://github.com/veneres/ltr-emb-analysis

**Table 1: Results in terms of BAcc and Acc for cross-validation on the development (d) and test sets (t).**

| Model | Distance | BAcc (d) | BAcc (t) | Acc (t) |
|---|---|---|---|---|
| MonoBERT | *cosine* | 0.57 ± 0.02 | 0.60 | 0.74 |
| | *l2* | 0.55 ± 0.02 | 0.55 | 0.76 |
| | *JSD* | 0.52 ± 0.02 | 0.57 | 0.77 |
| MonoRoBERTa | *cosine* | 0.75 ± 0.02 | 0.84 | 0.85 |
| | *l2* | 0.78 ± 0.02 | 0.81 | 0.85 |
| | *JSD* | 0.64 ± 0.01 | 0.66 | 0.79 |
| MonoELECTRA | *cosine* | 0.63 ± 0.02 | 0.64 | 0.75 |
| | *l2* | 0.66 ± 0.01 | 0.60 | 0.76 |
| | *JSD* | 0.50 ± 0.00 | 0.52 | 0.73 |

behavior starts from the very beginning ($T_0$) for certain metrics (e.g., *min-cosine*); this fact can have a huge impact also on the creation of new early-exit strategies since the inference could be stopped when $S_q$ and $S_d$ are too far away and the document scored is likely to be not relevant. An illustrative example of this behavior is depicted in Figure 2: here, we show two pairs of query and document token PDFs, each representing a score level (low/high). Interestingly, the closer the two distributions, the higher the score.

## 4 TOWARDS UNDERSTANDING MODEL PREDICTION ACCURACY

In our exploratory analysis, we found that the score predicted by the model is more aligned with the real relevance of a document when the query and document tokens follow a well-specified transformation in the space related to RQ1. Thus, to answer RQ2, we create different classifiers (shallow decision trees) using the information of similarity between $S_q$ and $S_d$ to predict the probability of having a misprediction, i.e., predicting that a document is relevant for the query when is not and vice-versa. That is, in the case of *cosine* and *l2*, we create a new dataset in which each row identifies a query/document pair and where it has as features all the different similarities for each block from $T_0$ to $T_{12}$, for a total of 52 features (using *min*, *max*, *avg*, and *center*), and 13 features for *JSD*. Since the goal of the prediction task is to predict the probability of having a misprediction, but the score predicted by the model is a real number between 0 and 1, we apply a min-max scaling to the score of the documents query-wise, and then we bin the score in 2 classes, low-score and high-score. Then, we binned the ground truth relevance in 2 classes: low-relevance within the interval $[0, 2]$ and high-relevance considering high-relevant only the documents with relevance 3. Other binning criteria could have been considered, however with these settings we want to emphasize in our misprediction classifier the role of very high-relevance judgments. Finally, let the binned score be $s_b$ and binned relevance be $r_b$, with value 0 when they represent a low score/relevance and value 1 when they represent a high score/relevance. The objective of our classifier is to predict the label $y = |r_b - s_b|$. It is easy to see that we have $y = 0$ when the model correctly scores the query/document pair, while we have $y = 1$ when the model assigns a high score to a non-relevant document or vice-versa. In Table 1, we present the accuracy obtained by our classifiers. The results are presented in terms of Balanced Accuracy (BAcc) and Accuracy (Acc) and AUC. We used BAcc in our classification problem to summarize the accuracy taking into account that the datasets are imbalanced and there are more non-relevant documents w.r.t. relevant documents [6]. For example, for MonoELECTRA (the most effective model) in *trec-dl-2019*, we have 2801 documents misclassified ($y = 1$) and 6459 correctly classified ($y = 0$). We used as training set and development set for our shallow decision trees the *trec-dl-2019* subset of MS MARCO, and we fine-tuned our classifier w.r.t. the number of leaves needed between $\{4, 8, 16, 32\}$ using 5-fold cross-validation. We then evaluate the classification accuracy on datasets created from *trec-dl-2019*. The results for BAcc and Acc are summarized in Table 1, while in Figure 3 we present the results for the AUC. From the results, we see we can achieve good classification performance with both the methods and distance measures, except for the *JSD* in MonoELECTRA where we get 0.52, close to random noise (our baseline). In addition, no distance seems to outperform the others in all the cases, and thus different metrics seem to be suitable for different models. Finally, we highlight that by inspecting the feature importance offered by the decision trees created, we can see that the most important ones are the ones in the last transformer blocks and in the first transformer blocks. For example, for MonoBERT and *cosine* distance, we have that the most important feature is the minimum distance at $T_{12}$ and the second most important feature is the distance at $T_1$. This furthermore proves our claim that considering the evolution of the embeddings along the transformer blocks is useful to understand the prediction made by the model.

## 5 CONCLUSIONS AND FUTURE WORK

In this work, we presented an initial analysis of the behaviors of the embeddings in Mono-X models during inference time. We propose two different research questions with the goal of zooming in on the latent space of the embeddings and understanding if there is a connection between the arrangement of the tokens, their score, and the prediction accuracy. Multiple future works can be based on the observations of this work, including the development of a new early-exit strategy observing the evolution of the tokens, improving the accuracy of the model by looking at the distribution of the tokens, creating new QPP methods and applying the same analysis also to dual encoder approaches. To conclude our work, given the analysis presented, we claim that we can answer the question: "Can Embeddings Analysis Explain Large Language Model Ranking?" in an affirmative way. In particular, we highlight that analyzing the distance between query and document tokens can be useful in understanding possible pitfalls of the model.

## ACKNOWLEDGMENTS

This work was partially supported by the SERICS project under the NRRP M4C2 Inv.1.3 PE00000014, by the iNEST project under the NRRP M4C2 Inv.1.5 ECS00000043 funded by the EU - NGEU, by the PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" funded by the European Commission under the NextGeneration EU programme, and by the Horizon Europe RIA "Extreme Food Risk Analytics" (EFRA), grant agreement n. 101093026.

# REFERENCES


[1] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, Addis Ababa, Ethiopia. https://openreview.net/forum?id=r1xMH1BtvB

[2] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 605–618. https://doi.org/10.1007/978-3-030-45439-5_40

[3] Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 55–65. https://doi.org/10.18653/v1/D19-1006

[4] Zeon Trevor Fernando, Jaspreet Singh, and Avishek Anand. 2019. A study on the Interpretability of Neural Retrieval Models using DeepSHAP. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1005–1008. https://doi.org/10.1145/3331184.3331312

[5] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556. https://doi.org/10.18653/v1/N19-1357

[6] John D Kelleher, Brian Mac Namee, and Aoife D'arcy. 2020. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT press.

[7] Minghan Li, Xinyu Zhang, Ji Xin, Hongyang Zhang, and Jimmy Lin. 2022. Certified Error Control of Candidate Set Pruning for Two-Stage Relevance Ranking. http://arxiv.org/abs/2205.09638 arXiv:2205.09638 [cs].

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. https://doi.org/10.48550/arXiv.1907.11692 arXiv:1907.11692 [cs].

[9] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2429–2436. https://doi.org/10.1145/3404835.3463254

[10] 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. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org, Barcelona, Spain. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[11] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* 452 (Sept. 2021), 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

[12] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. https://doi.org/10.48550/arXiv.1910.14424 arXiv:1910.14424 [cs].

[13] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. 2019. *Understanding the Behaviors of BERT in Ranking*. Technical Report arXiv:1904.07531. arXiv. https://doi.org/10.48550/arXiv.1904.07531 arXiv:1904.07531 [cs] type: article.

[14] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and Measuring the Geometry of BERT. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Vancouver, Canada. https://proceedings.neurips.cc/paper_files/paper/2019/hash/159c1ffe5b61b41b3c4d8f4c2150f6c4-Abstract.html

[15] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. In *Advances in Information Retrieval*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Vol. 11437. Springer International Publishing, Cham, 489–503. https://doi.org/10.1007/978-3-030-15712-8_32 Series Title: Lecture Notes in Computer Science.

[16] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951. https://doi.org/10.18653/v1/P19-1282

[17] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable Search Using Local Model Agnostic Interpretability. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, Melbourne VIC Australia, 770–773. https://doi.org/10.1145/3289600.3290620 20 citations (Crossref) [2022-02-08].

[18] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 618–628. https://doi.org/10.1145/3351095.3375234 12 citations (Semantic Scholar/DOI) [2021-11-04] 2 citations (Crossref) [2021-11-04].

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[20] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1281–1284. https://doi.org/10.1145/3331184.3331377

[21] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[22] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. (June 2021), 1–4. https://doi.org/10.18653/v1/2021.naacl-tutorials.1

[23] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. An Analysis of BERT in Document Ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1941–1944. https://doi.org/10.1145/3397271.3401325

[24] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query Understanding via Intent Description Generation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 1823–1832. https://doi.org/10.1145/3340531.3411999