Causal Probing for Dual Encoders

Wallat, Jonas; Hinrichs, Hauke; Anand, Avishek

DOI
10.1145/3627673.3679556

Publication date
2024

Document Version
Final published version

Published in
CIKM 2024 - Proceedings of the 33rd ACM International Conference on Information and Knowledge Management

Causal Probing for Dual Encoders

Jonas Wallat
L3S Research Center
Hannover, Germany
wallat@l3s.de

Hauke Hinrichs
L3S Research Center
Hannover, Germany
hinrichs@l3s.de

Avishek Anand
Department of Software Technology
Delft University of Technology
Delft, The Netherlands
avishek.anand@tudelft.nl

Abstract

Dual encoders are highly effective and widely deployed in the retrieval phase for passage and document ranking, question answering, or retrieval-augmented generation (RAG) setups. Most dual-encoder models use transformer models like BERT to map input queries and output targets to a common vector space encoding the semantic similarity. Despite their prevalence and impressive performance, little is known about the inner workings of dense encoders for retrieval. We investigate neural retrievers using the probing paradigm to identify well-understood IR properties that causally result in ranking performance. Unlike existing works that have probed cross-encoders to show query-document interactions, we provide a principled approach to probe dual-encoders. Importantly, we employ causal probing to avoid correlation effects that might be artefacts of vanilla probing. We conduct extensive experiments on one such dual encoder (TCT-ColBERT) to check for the *existence* and *relevance* of six properties: term importance, lexical matching (BM25), semantic matching, question classification, and the two linguistic properties of named entity recognition and coreference resolution. Our layer-wise analysis shows important differences between re-rankers and dual encoders, establishing which tasks are not only *understood* by the model but also *used* for inference.

CCS Concepts

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

Keywords

Information Retrieval, Interpretability, Language Models, Probing

ACM Reference Format:
Jonas Wallat, Hauke Hinrichs, and Avishek Anand. 2024. Causal Probing for Dual Encoders. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3627673.3679556

1 Introduction

Dual-encoder models are becoming increasingly common in modern information retrieval pipelines for the initial retrieval phase [20, 27, 66]. These encoders separately encode queries and documents into a joint embedding space where relevant queries and documents are represented in each other's proximity under a simple and efficiently computable distance function. With the improvement in efficient data structures for approximated nearest neighbor search [8, 26, 61], they are now popular choices for fast first-stage retrieval in retrieval augmented models among many other general applications like question answering, fact-checking, etc. Along with efficiency, dual-encoder models have achieved impressive results on several information retrieval benchmarks. In spite of the retrieval performance benefits, little is understood about the mechanisms they use to perform search and retrieval tasks – *to what degree do they understand classical notions of term-matching models? Do they identify entities in documents and queries for determining relevance? Do they perform basic NLP operations like co-reference resolution and entity linking, improving over term-matching, for determining relevance? Do they internally perform query classification?* Yet, a better understanding of how retrievers work can help identify failure cases [41] and allow for more effective training [3, 12, 64].

Consequently, a lot of the recent work on explainable information retrieval has focused on studying these models from various aspects [55, 57, 67, 69]. However, most of the popular and common techniques focus on explaining single decisions – a query-decision prediction (pointwise), a preference pair (pairwise), or an entire ranking for a query (listwise) [2]. Little work on model and dataset level improves our understanding of the general model capabilities, with the notable exceptions being [49, 64]. This paper aims to extend our understanding of BERT-based dual encoder models by using a family of techniques called model-probing that tries to understand to what extent a trained model exhibits well-understood linguistic and retrieval properties [58].

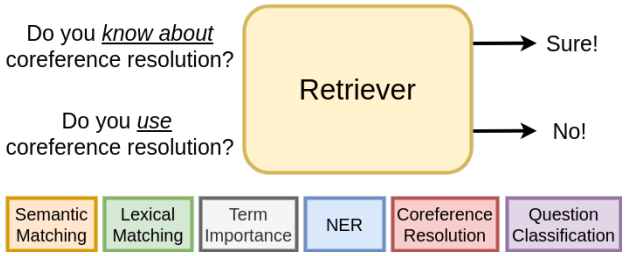


Figure 1: We causally probe dual encoders for retrieval for multiple IR and NLP abilities to understand which are relevant for the retrieval task. Specifically, we investigate the difference between probing (*presence* of information) and causal probing (*usage* of information in downstream tasks) for six IR abilities.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679556

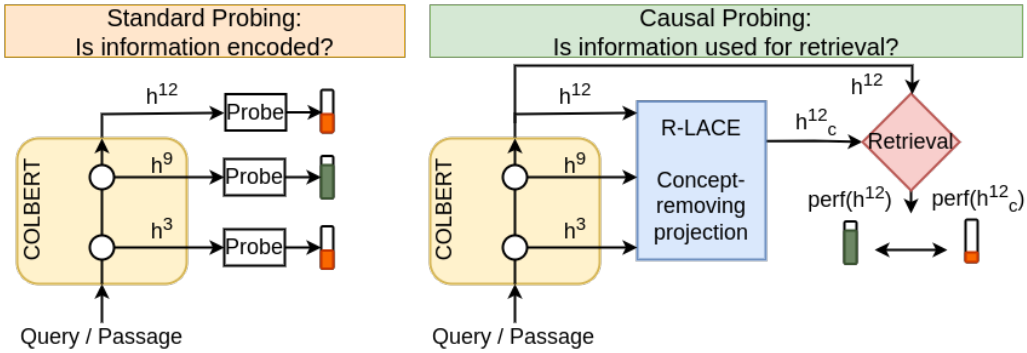


Figure 2: We use a combination of the standard probing setup to *locate* IR ability information in our dual encoder and causal probing to understand whether that information *is used* during retrieval.

Probing has been proposed as a method to analyze document representations by developing small probe tasks to characterize the ranking abilities of text encoders [34, 64, 65]. Instead of the commonly used local interpretability approaches that try to explain a certain decision [48, 56]; probing focuses on a holistic analysis of grounding the abilities of an already trained model to well-understood IR properties and abilities.

Probing tasks are usually framed as classification or regression problems, where the target variable relates to a specific linguistic or, in our case, an IR property of interest. The idea is to train a classifier (often referred to as the "probe") on top of the fixed representations produced by a task fine-tuned model, e.g., monoBERT [38], dual-encoders for text retrieval and ranking. If the probe can predict a certain property with high accuracy, it suggests that the information about that property exists in the representation. In this work, we are interested in probing fine-tuned text retrievers to check for the presence and utilization of IR and linguistic properties/abilities like matching, semantic similarity, named-entity recognition, etc.

Recent work has shown some innate limitations of the probing paradigm. Specifically, successful probing does not necessarily imply that the model uses that information for its primary task [46]. In other words, existing probing methods might show that a model's representations encode a certain IR or linguistic property, but these properties might not be used for the final task performance—text retrieval in our case. In this work, we intend to fill this gap in IR research by performing *causal probing* [44] to conclusively establish if a certain IR property is firstly *exhibited* in the document representations and secondly *utilized* in text retrieval (Figure 1).

1.1 Contributions

In this paper, we propose a method of causally probing all layers of dual encoders for retrieval to study the importance of several IR abilities. Besides our findings, we offer several methodical contributions. We apply causal probing to dual encoders and, to the best of our knowledge, are the first work of probing dual encoders on but the last layer. We detail our experimental setup in Chapter 3. Further, we analyze the otherwise neglected class of retrieval models. We consider several established IR abilities - lexical and semantic matching, question understanding, named entity recognition, and coreference resolution. We measure the relevance of such IR abilities by constructing counterfactual embeddings that do not contain these abilities and evaluating the retrieval performance when using the counterfactual embeddings. Figure 2 shows an overview of our approach. To the best of our knowledge, we are the first to apply such layer-wise analysis to dual encoders and the first to apply causal probing to IR models in general. Our experiments emphasize the relevance of BM25, RSJ term importance, NER, and question understanding for retrieval. Interestingly, removing the properties from the last layer is less critical than removing them from layers 8-11, suggesting that these layers contain much relevant information for the retrieval task. However, we observe no substantial impact of removing semantic similarity or coreference information. The code is available[1].

2 Related Work

Causal probing is part of a larger subfield of interpretability research called *mechanistic interpretability*, which focuses on the mechanisms by which models perform their tasks. There are several mechanistic studies in transformer-based (large) language models - for example, identifying functions of individual attention heads [21, 35, 53] or deciphering the usage of feed-forward layers in factual recall [19, 36]. The goal of the study is similarly to understand the processes that allow dense retrievers, specifically TCT-ColBERT [31], to perform retrieval. However, we do not focus on localizing or understanding individual attention heads, but investigate which layers and IR abilities are required to do so.

2.1 The Probing Paradigm

Probing was introduced by Conneau et al. [10] to analyze BERT [13] representations for lexical properties. Many studies have been conducted to analyze if text representations *learn* low-level syntactic features to high-level factual knowledge [58, 59]. In parallel to investigating linguistic, knowledge-based abilities, probing best practices have also emerged that attempt to answer the question – what are the best practices to probe effectively? [24, 39, 40]. Most notably, Ravichander et al. [46] found that probe classifier would also achieve high accuracy on tasks not related to the downstream

[1] https://github.com/Heyjuke58/causal_probing

task and, therefore, questioned whether probing accuracy is answering the implicit question of relevancy towards to task. To counteract this problem, both Pimentel and Cotterell [39] and Voita and Titov [62] proposed approaches to measure the ease of extraction together with the accuracy of the task to understand better how usable the property is for the model. Lately, several works moved from probing layers to probing individual neurons. This is either utilized to investigate factual information in neurons [11] or as a general methodology to investigate how localized information is encoded [22]. Additional work has found that language models encode functions in their embedding spaces Hendel et al. [23], Todd et al. [60], further motivating whether IR abilities are encoded.

Yet, much of the work comes short of the implicit goal of probing studies. We are less interested in whether some property (e.g., subject-verb agreement) can be decoded from the model's embeddings but rather in whether this information is used for inference.

2.2 Causal Probing

Since the standard probing setup is limited to identifying whether information can be decoded from embeddings, causal probing methods have been developed to investigate the abilities' *relevance for downstream tasks* ([4, 14, 46] inter alia). This is usually done by constructing counterfactual embeddings that do not contain certain information and then testing the impact on a downstream task by comparing the performance of normal embeddings to the counterfactual ones. Elazar et al. [14] first suggested this approach and investigated the impact of part-of-speech information on the masked language modeling task. Similarly, Lasri et al. [29] investigate the use of grammatical number information in BERT, finding different encodings for verbs and nouns that are used for language modeling. Lastly, Rozanova et al. [52] use the causal probing approach and investigate natural language inference (NLI), but other than previous studies build counterfactual representations by removing everything except the task-specific information, finding that it better aligns with theoretical expectations in the NLI case.

Central to the causal probing approach is building counterfactual representations (i.e., representations without the specific property under investigation). Previous work has focused on iterative nullspace projection (INLP) [43] - a method in which linear classifiers are iteratively trained to predict the property (e.g., gender bias), and the information used by these classifiers is iteratively guarded by projecting the input to the nullspace of the classifier. This was later refined using a minimax optimization problem formulation [45]. Our work builds on the work of Ravfogel et al. [45] that is theoretically well-founded.

2.3 Analysis of IR Models

In this section, we review the recent works using the probing paradigm or related approaches to shed light on mechanisms and learned information of IR models. Zhan et al. [68] investigate the attention patterns of ranking models and find that large amounts of the attention are offloaded to punctuation and other low-information tokens. Choi et al. [9] analyze the attention maps of ranking models and find these to contain inverse document frequency information. Another line of interpretability works aims to explain ranking models with understandable concepts. Several works either tested whether the model's predictions agree with IR axioms [6, 49] or tried explaining predictions by aligning them to such [63]. Sen et al. [54] use a similar approach but utilize a linear classifier with the coefficients of term frequency, document frequency, and document length to approximate the model's predictions. Adolphs et al. [1] employ query embeddings to generate query reformulations and show that these embeddings can be moved in latent space to retrieve relevant paragraphs. Following the finding that embeddings can be projected into the vocabulary space, often resulting in understandable concepts [18], Ram et al. [42] investigate dual encoder representations and failure-cases in a similar manner. Relatedly, Liu and Mao [32] project representations of multi-vector dense retrievers into the vocabulary space and find that different vectors can address different information needs from passages.

Several probing studies have analyzed ranking models for a wide variety of NLP and IR-related tasks; Fan et al. [15] probed different IR models for 16 linguistic tasks (lexical, syntactic, and semantic), such as part-of-speech or polysemy. MacAveney et al. [34] probed a large set of ranking and retrieval models on three categories of tasks: matching abilities, sensitivity to manipulation, and sensitivity to writing styles. Further, Lovón-Melgarejo et al. [33] probed language models for hierarchical properties, finding their injection can improve LM's understanding of hierarchy. Yet, these studies were limited to analyzing the probing performance on the models' last layer. Wallat et al. [64] used probing to identify layers that contain the most task-specific information and applied this knowledge to design a multi-task learning setup to train better ranking models. Chen et al. [7] use causal interventions to reverse engineer the relevance judgement of a dense retriever model, identifying that a group of attention heads adhere to term-frequency axioms.

In a behavioural study, Formal et al. [16] investigate the matching abilities of ColBERT [28], a common dual encoder model. By analyzing exact and soft matches w.r.t. term importance, Formal et al. find that the model captures a notion of term importance and seems to rely on that information for identifying important terms. In a follow-up work, Formal et al. [17] find that the inability to identify important terms in unseen distributions is one reason for the poor generalization abilities of ranking and retrieval models.

This study aims to ground retrieval performance to well-understood IR properties, making dense retrievers more understandable. While related to existing work, our work employs the *causal probing* approach to understand matching and other abilities of retrieval models not only on the last but on all layers.

3 Probing Bi-Encoder Ranking Models

3.1 Preliminaries: Causal Probing

When analyzing a dual encoder, our goals are two-fold. By *probing* the model, we first want to show that an ability is *exhibited*. In a second *causal probing* step, we study task relevancy by investigating the impact on retrieval performance of surgically removing IR abilities. Probing entails training a small classifier f to predict a set of ranking abilities A (e.g., `BM25` scores) from the embeddings of a ranker Φ. To do so, we construct training and test data sets $\mathbf{A}$ where the input is query and passage pairs, and the target is specified by the ability (e.g., a `BM25` score). We then train the classifier (referred to as the "probe") on fixed representations of our model to predict

Algorithm 1: Property removal using R-Lace

Input: Data (X, y), Loss ℓ, projection rank k, outerloop T, inner loop M
Output: A projection matrix P that neutralizes a rank space

```
// Initialization
initialize predictor θ ∈ R^D randomly
initialize predictor P ∈ R^{DXD} randomly
for i = 1 to T do
    for j = 1 to M do
        θ ← SGDUpdate (∂ℓ(y, XPθ)/∂θ);
    end
    for j = 1 to M do
        P ← SGDUpdate (∂ℓ(y, XPθ)/∂P)
        P ← ½(P + P^T) // Ensure P is symmetric

        P ← FantopeProjection(P, k) // Project on the
            fantome

    end
end
// Recomputing P
U, D = spectralDecomposition (P) // Perform SVD that
    reduces the rank by k

P ← U[: −k, :]^⊤ U[: −k, :]
return P
```

Algorithm 2: Causal Probing Dual Encoders

Input: Task dataset $\mathbf{a} = \{(q, p, y), ...\}$, Ranker Φ
Output: Causal probing results

```
for l in layers(Φ) do
    Initialize empty list: embeddings;
    for (query, passage, y) in a do
        queries ← avg_pool(φ^l(q));
        passages ← avg_pool(φ^l(p));
        embeddings ← avg_pool(queries, passages);
    end
    // Get probing performance on task a
    acc(f(embeddings, y));
    // Get concept-removing projection
    if task a is a classification task then
        P ← LACE(φ^l(a), y);
    end
    else if task a is a regression task then
        P ← R-Lace(φ^l(a), y, rank);
    end
    // Check whether the concept-removing
       projection is able to remove the ability a
    assert acc(f(P embeddings, y)) == majority;
    // Evaluate retrieval performance on TREC-DL
    ndcg ← Φ(TREC);
    // Evaluate retrieval performance on TREC-DL
       after removing information from layer l
    ndcg_c ← Φ(TREC) where φ^l_c = P φ^l;
end
```

the targets. If the classifier's performance is above chance, Φ is said to exhibit that ability. Since this test can establish the presence of an ability but not its usage, we apply causal probing methods in a second step to address this limitation [14, 29]. Causal probing measures the relevance of an ability a by removing it from Φ's embeddings and then evaluating the impact on the total retrieval performance. If Φ uses a for retrieval, we would observe reduced performance after a's removal.

Problem Statement. Given a document, a query encoder Φ, and an ability a, we want to determine a's causal effect on the retrieval performance.

3.2 Property Removal

To causally determine if a certain ability a is responsible for retrieval task performance, we posit that the *removal of the ability* would result in the reduction of task performance. Specifically, we would want to remove the ability a from the input representation space ϕ induced by the ranker Φ so that the output counterfactual representation space ϕ_c does not encode a. Note that the ability to encode a is determined by the performance on the probe task. We use a framework [45] that uses linear projections for concept removal in the context of bias detection and removal. Mathematically, we want to find the projection matrix P, such that counterfactual embeddings $\phi_c = P\phi$ cannot be used to classify items in the probe task a. For regression abilities, we use an analytical formulation of such a projection given by

$$P = I - \frac{X^\top \boldsymbol{y}\boldsymbol{y} X}{\boldsymbol{y}^\top X X^\top \boldsymbol{y}} \tag{1}$$

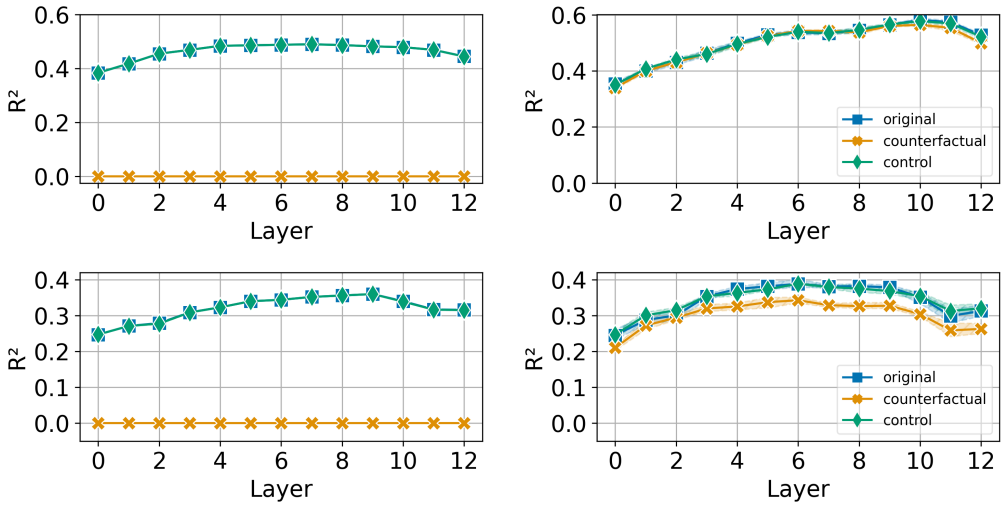
The iterative and relaxed version (R-Lace), which we apply to classification tasks, is given by a minimax game. Generally, minimax games are hard to optimize, except for the group of convex-concave games where the inner optimization is convex and the outer concave. However, R-Lace solves this by relaxing the only source on non-convexity (the set of potential orthogonal projection matrices $\mathcal{P}_k$) to its convex hull (which in this case is the fantope [5]).

$$\mathcal{F}_k = conv(\mathcal{P}_k) \tag{2}$$

The minimax game is then given by

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^D} \max_{P\in\mathcal{F}_k} \sum_{n=1}^{N} \ell\Big(y_n,\ g^{-1}\Big(\boldsymbol{\theta}^\top P \boldsymbol{x_n}\Big)\Big) \tag{3}$$

where θ is the classifier's parameter space and $\ell(\cdot,\cdot), g^{-1}(\cdot,\cdot)$ are the loss function and the activation function respectively. Optimization is then achieved by alternating the optimization steps of the outer and inner optimization problems while holding the other one fixed. Algorithm 1 depicts how R-Lace can be used to get a property-removing projection P. Note that with property removal, the aim



(a) Linear probe **(b) Non-linear probe**

Figure 3: Probing and property removal results for TI (top) and BM25 (bottom).

is to be minimally invasive and remove as little information not related to a as possible. In the analytical solution, this is given by a rank 1 projection P. For the relaxed R-LACE case, the subspace rank is a hyperparameter, and we experiment with different ranks of P given that a rank of 1 was insufficient and report the findings.

3.3 Causal Probing for Dual Encoders

Most of the work for probing neural rankers has been performed only for representation space induced by joint query-document representation [64]. Therefore, the investigations are limited mostly to the re-ranking phase. We, however, are the first to layer-wise probe (causal and otherwise) dual encoder models.

We utilize the recent dual encoder TCT-COLBERT [31] as our subject model Φ. TCT-COLBERT is a 12-layer dense dual-encoder retriever model based on the BERT [13] transformer architecture. Other than contextual models, it independently encodes query and passage, resulting in two separate embeddings. To score the relevance of a passage w.r.t. a query, TCT-COLBERT uses dot-products between the corresponding embeddings.

We then construct task datasets $\mathbf{a}$ for all abilities a in our list of IR abilities A (c.f. Section 3.4). Examples of $\mathbf{a}$ for a task a contain a query q, a passage p, a target value y, and optionally spans sp of the position of specific tokens in s.

Probing contextual encoders is straightforward since there is only one contextual embedding containing information from both query and passage. Probing dual encoders, however, requires a different strategy since we obtain individual embeddings for query and passage that only interact after the final layer. Since many of our tasks, such as semantic similarity, require that interaction, we propose the following strategy; For $q, p, y \in \mathbf{a}$ and a given layer l, we retrieve Φ^l's token embeddings and use average pooling to get one single vector for the query and a single vector for the passage. These vectors are then pooled again and used as input to the probe classifier, together with the corresponding label y.

Causal probing requires obtaining the *concept-nullifying projection matrix* P using (R)-LACE. Depending on the task type, we use the analytical solution (LACE, for regression) or the relaxed version (R-LACE, for classification). Additionally, we distinguish between token or span-level (such as NER) and sequence-level tasks (such as semantic similarity). While we use individual tokens or spans to compute the (R)-LACE projection in the former case, we pool over the token embeddings for sequence-level tasks. The resulting projection is then applied to the entire sequence of tokens (both query and passage tokens). Algorithm 2 shows this process.

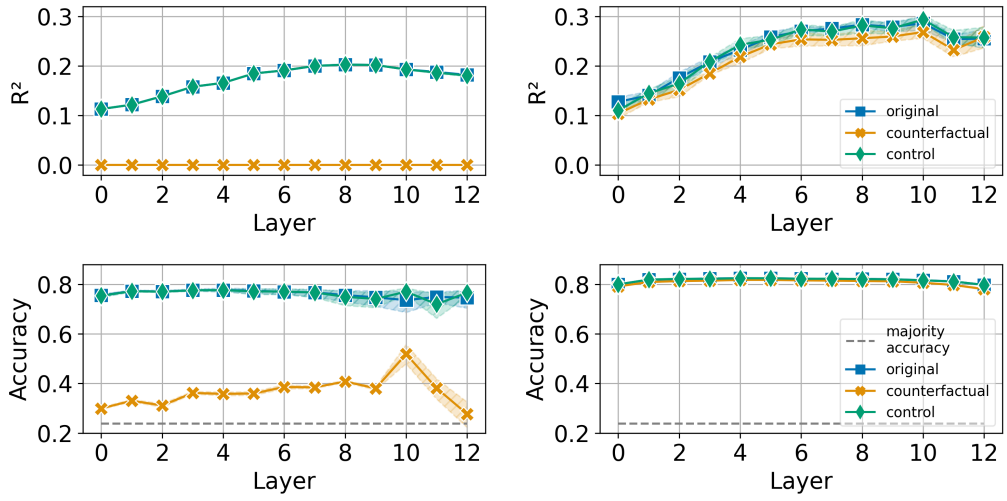
3.4 IR Abilities

We utilize a selection of IR abilities that are strongly grounded in the ranking and retrieval literature. If not noted otherwise, we build the probe datasets by generating task-specific labels for 60k randomly sampled query-passage pairs from MS MARCO [37]. This is similar to existing work [64], from which we use the existing datasets for BM25, NER, and coreference resolution. We then extend these with additional tasks and apply them in our different settings.

BM25. As one way of measuring lexical similarity, this dataset is constructed by using the BM25 algorithm [51] to produce BM25 scores between query and passage. To correctly predict the BM25 score, the model has to compute IDF values, as well as be aware of the average document lengths in the corpus.

Term Importance or TI. Similar to Formal et al. [17], we investigate the matching abilities of IR models by inspecting the model's ability to understand the Robertson-Spärck-Jones (RSJ) weight [50]. These RSJ weights measure the term importance of a token w.r.t. a query and a corpus. It is computed as follows:

$$RSJ(t, q, C) = \log \frac{p(t|\mathcal{R})p(\neg t|\neg\mathcal{R})}{p(\neg t|\mathcal{R})p(t|\neg\mathcal{R})} \tag{4}$$



(a) Linear probe (b) Non-linear probe

Figure 4: Probing and property removal results for SEM (top) and NER (bottom).

Semantic Similarity or `Sem`. Given that semantic matching has been one of the main improvements of the embedding-based machine learning model, we test for TCT-ColBERT's usage of semantic matching to retrieve passages. We generate labels for the semantic similarity task using Sentence-Transformer [47].

Named Entity Recognition or `NER`. Since many queries evolve around entities, the ability to detect such could be relevant. Therefore, we tested to what extent this ability affects the overall retrieval performance. This dataset is created using the Spacy [25] named entity recognizer to find entities in the passages.

Coreference Resolution or `Coref`. Related is the ability to match surface forms to one entity. To understand to what degree this is used by TCT-ColBERT, we additionally test this ability. Usually, this dataset would require matching an entity phenotype from the query to another mention in the passage (as done in [64]). Since we are using a dual encoder in this study and the query and document are processed independently, we construct this dataset by finding coreference examples only in the passage. Analogously, we also train the R-Lace projection only on `Coref` examples in the passage.

Question Classification or `QC`. We believe one central ability to find the correct information is to understand what kind of query is given. Thus, we include question type classification as a task. We use the dataset provided by Li and Roth [30] containing 5453 questions and use the coarse labels (abbreviation, entity, description, human, location, numeric value). Given that `QC` is only defined for the query, we train the R-Lace projection only on the query and apply the resulting projection only to the query embeddings.

3.5 Probing Metrics

For our classical probing experiments, we report the classifier's accuracy and, in the regression case, the regressor's coefficient of determination (R^2). We use R^2 over mean squared error (MSE) as R^2 is scaled so that comparisons between datasets become easier.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

where $SS_{res} = \sum_i e_i^2$ is the sum of squares of residuals and $SS_{tot} = \sum_i (\overline{y} - y_i)^2$. The residual is $e_i = y_i - f_i$ and $\overline{y}$ is the mean of the targets. An R^2 value of 1 would denote the regression model to perfectly fit the data.

4 Do Dual-Encoder Representations contain our IR Abilities?

As a first step, we want to understand whether the abilities are actually *encoded* by the model's embeddings and *whether we can remove* the information using (R)-Lace. We consider probing a fine-tuned dual encoder [31] for the IR abilities in Section 3.4. We utilize the layer-wise probing procedure detailed in Section 3.3. Given that Lace will remove linear-encoded information from the embeddings, we train linear probe models to predict the probe task (original) and compare it with three baselines; First, the accuracy achieved by a majority classifier. This is to understand whether the property is encoded in the embeddings in the first place (majority). Second, the performance of the probe model on the counterfactual embeddings ϕ_c that result from Lace removing the (linear) probe task information from the original embeddings. This is a test of whether Lace can remove all of the ability-related information, and we expect the probe classifier to perform worse (and at or below majority) for the counterfactual embeddings. Lastly, we remove random dimensions with a similar rank as those removed by Lace (control) to understand if the reduced performance is caused by Lace selectively removing probe task information or destroying the embeddings. Ideally, the control performance should be close to the original performance. Every run is repeated five times using the embeddings of every layer. We also include the same experiments

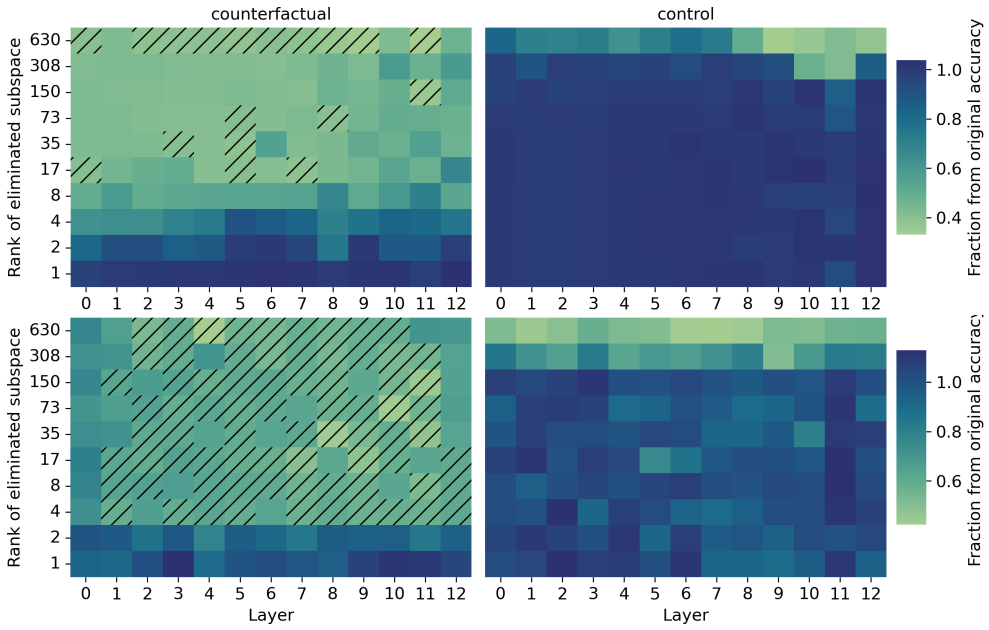


Figure 5: Top row: NER. Bottom row: QC. Accuracy decreases after applying the R-Lace counterfactual projection (left) or control projection (right) at each layer with increasing rank of the eliminated subspace. A value of 1 indicates no accuracy decrease. The pattern indicates where the accuracy is equal to or less than the majority accuracy (0.24 for NER, 0.23 for QC).

with an MLP probe model. By comparing the performance of linear and non-linear probes, we aim to understand what amount of task information might be encoded in a non-linear fashion. Given that (R)-Lace is only able to remove linear information, this yields additional contextualization.

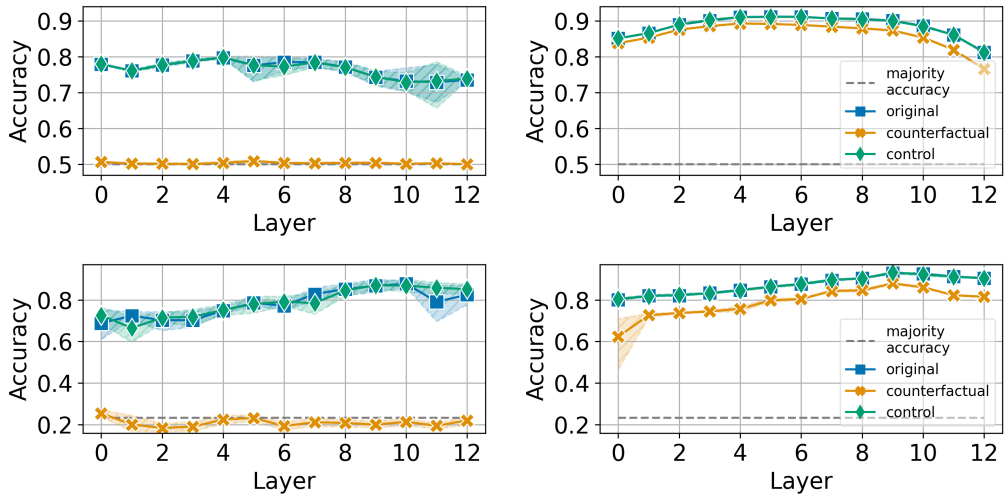
The results for the TI probe task are shown in Figure 3. Despite slight differences between the linear and the MLP probe models, we observe similar trends and performance; the performance increases up to layers 7-9 and then slightly decreases toward the final layer. We note that Lace completely removes the linear information. As expected, we only observe slight decreases in MLP performance after removing TI information using Lace. Additionally, removing TI information using Lace seems to be sufficiently selective as we do not see reduced performance with our control.

Similar trends can be seen for the other two regression tasks of **BM25** (Figure 3) and **Sem** (Figure 4, top), where both linear and MLP probe models peak at layer 9 and 8 respectively and slightly decrease in performance toward the final layer. Similar to the TI results, we observe Lace to remove all linear information regarding BM25 and Sem.

Next, we will discuss the three classification tasks of NER, Coref, and QC for which we used the relaxed version (R-Lace) to construct the counterfactual embeddings. While the analytical Lace projection always removes a minimal subspace of rank 1, the size of the subspace becomes a hyperparameter in R-Lace.

What is the right subspace size to remove? First and foremost, we observe that removing a subspace of rank 1 is enough to remove all linear information in the Coref case (c.f. Figure 6, top left), but not for NER and QC. We, therefore, remove subspaces of increasing rank from the model's embeddings using R-Lace. Since removing more and more information from the embeddings comes at the risk of destroying the embeddings, we similarly remove random subspaces of increasing rank from the embeddings to understand the impact on embedding integrity. The results are depicted in Figure 5. We observe that in the case of QC, using R-Lace to remove a subspace of rank 4 reduces the probing performance to the majority for most layers with non-zero but also non-substantial impact on the embeddings (control). For NER, a very high-ranking subspace would need to be removed to delete (almost) all NER information from the embeddings. However, removing such a big subspace comes at the cost of destroying other important information from the embedding. We, therefore, select a rank for 8 for the NER experiments since it removes a considerable amount of NER information without impacting other information in the embeddings.

The results for the **NER** task are shown in Figure 4 (bottom). Other than for the regression tasks, the NER performance is rather stagnant over the layers both for the linear and the MLP probe model. R-Lace with a subspace rank of 8 is able to remove significant parts of the NER information but does not entirely reach majority accuracy. For **Coref** (see Figure 6, top), we observe quite a difference in linear and MLP probe performance. While both probes peak in performance in the middle layers (4 and 5) and then decrease, as observed with other probe tasks, there seems to be a larger discrepancy between the linear and MLP models. This might hint at not all Coref information being linearly encoded by the model. Yet, R-Lace removes all the linear Coref information from the embeddings. Lastly, the results for **QC** (Figure 6, bottom) locate most of the QC information around layer 10 with slight drops in performance toward layer 12. The linear QC information is successfully removed by R-Lace.



(a) Linear probe (b) Non-linear probe

Figure 6: Probing and property removal results for COREF (top) and QC (bottom).

4.1 Insights

Can we remove IR ability information? For the regression tasks TI, BM25, Sem, we utilize the closed-form solution (Lace) and observe perfect removal of IR subtask information from the model's embeddings. For two of our classification tasks (Coref and QC), we again observe close-to-perfect information removal using the iterative R-Lace. For the NER task, we find the R-Lace projection to remove much of the task information, but the classifier performance not quite dropping to majority accuracy.

Where in the model is the information located? Task-specific information is mostly located in layers 4 – 10, with slight decreases in task performance toward the final layer. This further motivates the layer-wise analysis of IR models.

5 Are the Abilities used for Ranking?

Now that we have established that 1) the model contains linear task information on our probe tasks and 2) we can successfully remove that task information using (R)-Lace, we can investigate the most important question: *Are these tasks relevant for retrieval?*. To answer this question, we use (R-)Lace to produce counterfactual embeddings as discussed in Section 4 and inject these at individual layers. We then use the last layer's embeddings on the retrieval task and report the impact of the intervention. The results for our six tasks and the baseline performance of the (unaltered) model are given in Figure 7. Lower NDCG scores at a given layer indicate an IR ability's usage for the retrieval task.

5.1 Results

Non-relevant Abilities. Judging by the impact of our causal analysis, we do not find evidence of either Sem or Coref being used during retrieval. The model seems to rely on lexical over semantic matching. Since Coref is also the task with the largest discrepancy

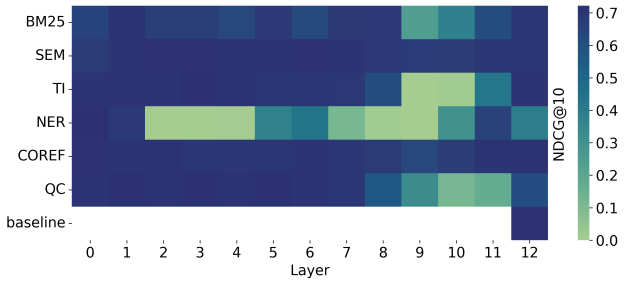


Figure 7: Impact of ability removal on retrieval performance.

between linear and non-linear information (c.f. Section 4), it might be that *linear* Coref information is not used for retrieval.

Relevant Abilities. We observe a negative impact on retrieval performance after removing the remaining properties, suggesting that these are all used to varying extents during retrieval. Lexical matching (measured by BM25 and TI) seems to be most relevant in layers 9 and 10. This finding coincides with our earlier analysis in Section 4, where we found these layers also to contain the most information regarding these abilities. Similarly, we find QC to be most utilized in the upper layers (around layers 10 and 11). Removing these abilities from other layers (especially the last layer) only has a small impact on retrieval quality. This would suggest that linear ability information is relevant at some point in TCT-ColBERT's embedding space but is then transformed or aggregated so that removing it directly from the last layer no longer has an effect. Solely NER is found to be relevant almost irrespectively of the layer under investigation.

Insight. The model utilizes linear information about all IR subtasks except for Sem and Coref for retrieval. While the most important layers vary slightly, they almost all center around the layers 9-11.

Study	Model Type	TI	BM25	SEM	NER	COREF	QC
[64]	Contextual (re-rankers)	/	$\uparrow, L_5$	$\uparrow, L_4$	$\downarrow, L_4$	$\uparrow, L_6$	/
[15]	Contextual (retrieval)	/	/	/	$\leftrightarrow, L$	$\leftrightarrow, L$	/
[34]	Dual (retrieval)	L	L	/	/	/	/
[16, 17]	Dual (retrieval)	R	/	/	/	/	/
[7]	Dual (retrieval) 6 layers	$R_{4,5}$		/	/	/	/
Ours	Dual (retrieval)	L_7, R_9	L_9, R_9	L_8	L_4, R_{many}	L_4	L_{10}, R_{10}

Table 1: Contextualization of our study with related studies. We denote studies that locate information at a specific layer with L_{layer} and use R_{layer} to denote our findings of the most relevant layer. Studies that only investigate the last layer are denoted with L or R without indices. For studies that compare the probing performance of the ranking model to pre-trained models, we use arrows to indicate whether the IR model outperforms the pre-trained model or not.

6 Discussion

How do our results relate to existing studies? To understand how the results of this causal probing study relate to existing research, we present the most relevant other studies in Table 1.

Several other studies identify IR ability information in the last layer of IR models [15, 34]. Yet, the models under investigation, tasks, and methodology vary between studies. Additionally, by limiting the probing approach to the last layer, it does not shed light on where in the model the information is *stored* or *used*. Given that we find all properties to peak in intermediate layers, this information is relevant to understanding the internal information processing of such ranking models (in agreement with [64]). Compared with Wallat et al. [64], who also probe layer-wise, we observe some disagreement in terms of the location of the knowledge and what tasks seem to be important. This might be due to multiple reasons; First, their work judges the relevance of IR subtasks by measuring how easily decodable - how *available* - the information is to the model and compares this with pre-trained BERT models. This ease of extraction might be a proxy but not a real test for whether the model *uses* the information. So even though `Sem` and `Coref` are easier to decode in their study, it does not guarantee the usage. Second, when comparing the layers in which information peaks, we find the contextual re-ranker [64] to peak in earlier layers. This might be a result of the different paradigms or contextual and dual encoders. In the latter, query and passage embeddings are only interacting after layer 12. We hypothesize that given many of our tasks only make sense in the interaction between query and passage (e.g., `Sem`), it makes sense for the model to preserve this information up until later layers. Chen et al. [7] checked whether individual attention heads adhere to a term frequency axiom (TFC1, "Prefer documents with more query term occurrences.") and found this information to have an effect on the relevance prediction of a 6-layer dual-encoder. Specifically, this information has been used by the model in layers 4 and 5. This is related to our IR subtasks of `TI` and `BM25`, which also have been observed to have an impact not in the ultimate but in the penultimate layers of the model.

Impact of our Results. The results of this offer a better understanding of dual encoder models for retrieval and the information they use to perform that task. Given the rise of retrieval-augmented generation (RAG), understanding the retrieval component is especially important. Causal analysis of IR models allows for answering the question of what information is being used for relevance estimation, which will hopefully spur further research. Especially, the findings that semantic similarity did not seem to be causally important for the retrieval task warrants more investigations. Further, we found that removing the IR abilities from the last layer did not substantially impact the model, posing the question of what information is being used in the last layer. Is it a composite of the individual abilities or some other property that we did not test for?

While the findings of this study are descriptive, they could be used to produce more effective IR models by informing more effective training setups on how to combine retrieval training signal with auxiliary information [3, 12, 64]. Lastly, more robust and fair models can be build by understanding which information is being used and, potentially, removing unwanted knowledge.

7 Conclusion

In this work, we investigate which subtasks are relevant for TCT-ColBERT to perform retrieval. To do so, we collect a selection of established IR abilities (term importance, BM25, semantic similarity, named entity recognition, coreference resolution, and question classification). We first show that information on all of the tasks above can be located in the model - and, further, can be removed from the embeddings using (R)-Lace. Using a layer-wise analysis, we show that most of the task-specific information seems to be encoded in the middle layers (4-10). In the second step, we investigate the importance of these tasks for retrieval by removing the task information from our model's embeddings. We remove task information from single layers at a time and report the impact on the final retrieval performance - where we observe high relevance of BM25, term importance, and question classification information in layers 9-11 and the widespread importance of named entity recognition. Surprisingly, semantic similarity and coreference resolution do not seem to be relevant for the retrieval task. To the best of our knowledge, this is the first work of causally probing IR models and investigating the importance of IR subtasks for *inference*.

Acknowledgments

This work was supported by the Lower Saxony Ministry of Science and Culture (MWK), in the zukunft.niedersachsen program of the Volkswagen Foundation (HybrInt).

References

[1] Leonard Adolphs, Michelle Chen Huebscher, Christian Buck, Sertan Girgin, Olivier Bachem, Massimiliano Ciaramita, and Thomas Hofmann. 2022. Decoding a Neural Retriever's Latent Space for Query Suggestion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 8786–8804. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.601

[2] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable Information Retrieval: A Survey. *CoRR* abs/2211.02405 (2022). https://doi.org/10.48550/arXiv.2211.02405 arXiv:2211.02405

[3] Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. 2023. Injecting the BM25 Score as Text Improves BERT-Based Re-rankers. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13980)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 66–83. https://doi.org/10.1007/978-3-031-28244-7_5

[4] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Comput. Linguistics* 48, 1 (2022), 207–219. https://doi.org/10.1162/coli_a_00422

[5] Stephen P. Boyd and Lieven Vandenberghe. 2014. *Convex Optimization*. Cambridge University Press. https://doi.org/10.1017/CBO9780511804441

[6] Arthur Câmara and Claudia Hauff. 2020. Diagnosing BERT with Retrieval Heuristics. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 12035)*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer, 605–618. https://doi.org/10.1007/978-3-030-45439-5_40

[7] Catherine Chen, Jack Merullo, and Carsten Eickhoff. [n. d.]. Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models. In *Proceedings of the 47rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2024, July 14–18, 2024, Washington, DC, USA*.

[8] Qi Chen, Bing Zhao, Haidong Wang, Mingqin Li, Chuanjie Liu, Zengzhong Li, Mao Yang, and Jingdong Wang. 2021. SPANN: Highly-efficient Billion-scale Approximate Nearest Neighborhood Search. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 5199–5212. https://proceedings.neurips.cc/paper/2021/hash/299dc35e747eb77177d9cea10a802da2-Abstract.html

[9] Jaekeol Choi, Euna Jung, Sungjun Lim, and Wonjong Rhee. 2022. Finding Inverse Document Frequency Information in BERT. *CoRR* abs/2202.12191 (2022). arXiv:2202.12191 https://arxiv.org/abs/2202.12191

[10] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 2126–2136. https://doi.org/10.18653/v1/P18-1198

[11] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 8493–8502. https://doi.org/10.18653/V1/2022.ACL-LONG.581

[12] Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W. Bruce Croft. 2017. Neural Ranking Models with Weak Supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, Noriko Kando, Tetsuya Sakai, Hideo Joho, Hang Li, Arjen P. de Vries, and Ryen W. White (Eds.). ACM, 65–74. https://doi.org/10.1145/3077136.3080832

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation With Amnesic Counterfactuals. *Trans. Assoc. Comput. Linguistics* 9 (2021), 160–175. https://doi.org/10.1162/tacl_a_00359

[15] Yixing Fan, Jiafeng Guo, Xinyu Ma, Ruqing Zhang, Yanyan Lan, and Xueqi Cheng. 2021. A Linguistic Study on Relevance Modeling in Information Retrieval. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 1053–1064. https://doi.org/10.1145/3442381.3450009

[16] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A White Box Analysis of ColBERT. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 12657)*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer, 257–263. https://doi.org/10.1007/978-3-030-72240-1_23

[17] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2022. Match Your Words! A Study of Lexical Matching in Neural Information Retrieval. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13186)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 120–127. https://doi.org/10.1007/978-3-030-99739-7_14

[18] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 30–45. https://doi.org/10.18653/V1/2022.EMNLP-MAIN.3

[19] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 5484–5495. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.446

[20] Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-End Retrieval in Continuous Space. *CoRR* abs/1811.08008 (2018). arXiv:1811.08008 http://arxiv.org/abs/1811.08008

[21] Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2023. Successor Heads: Recurring, Interpretable Attention Heads In The Wild. *CoRR* abs/2312.09230 (2023). https://doi.org/10.48550/ARXIV.2312.09230 arXiv:2312.09230

[22] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding Neurons in a Haystack: Case Studies with Sparse Probing. *CoRR* abs/2305.01610 (2023). https://doi.org/10.48550/ARXIV.2305.01610 arXiv:2305.01610

[23] Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 9318–9333. https://aclanthology.org/2023.findings-emnlp.624

[24] John Hewitt and Percy Liang. 2019. Designing and Interpreting Probes with Control Tasks. 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, 2733–2743. https://doi.org/10.18653/v1/D19-1275

[25] M Honnibal and I Montani. 2017. Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application. https://spacy. io* (2017).

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-Scale Similarity Search with GPUs. *IEEE Trans. Big Data* 7, 3 (2021), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[27] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 6769–6781. https://doi.org/10.18653/V1/2020.EMNLP-MAIN.550

[28] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 39–48. https://doi.org/10.1145/3397271.3401075

[29] Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the Usage of Grammatical Number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 8818–8831. https://doi.org/10.18653/v1/2022.acl-long.603

[30] Xin Li and Dan Roth. 2002. Learning Question Classifiers. In *19th International Conference on Computational Linguistics, COLING 2002, Howard International*

House and Academia Sinica, Taipei, Taiwan, August 24 - September 1, 2002. https://aclanthology.org/C02-1150/

[31] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP, RepL4NLP@ACL-IJCNLP 2021, Online, August 6, 2021*, Anna Rogers, Iacer Calixto, Ivan Vulic, Naomi Saphra, Nora Kassner, Oana-Maria Camburu, Trapit Bansal, and Vered Shwartz (Eds.). Association for Computational Linguistics, 163–173. https://doi.org/10.18653/V1/2021.REPL4NLP-1.17

[32] Qi Liu and Jiaxin Mao. 2023. Understanding the Multi-vector Dense Retrieval Models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos (Eds.). ACM, 4110–4114. https://doi.org/10.1145/3583780.3615282

[33] Jesús Lovón-Melgarejo, José G. Moreno, Romaric Besançon, Olivier Ferret, and Lynda Tamine. 2024. Probing Pretrained Language Models with Hierarchy Properties. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 14609)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 126–142. https://doi.org/10.1007/978-3-031-56060-6_9

[34] Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, and Arman Cohan. 2022. ABNIRML: Analyzing the Behavior of Neural IR Models. *Trans. Assoc. Comput. Linguistics* 10 (2022), 224–239. https://doi.org/10.1162/tacl_a_00457

[35] Callum McDougall, Arthur Conmy, Cody Rushing, Thomas McGrath, and Neel Nanda. 2023. Copy Suppression: Comprehensively Understanding an Attention Head. *CoRR* abs/2310.04625 (2023). https://doi.org/10.48550/ARXIV.2310.04625 arXiv:2310.04625

[36] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html

[37] 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. http://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[38] Rodrigo Frassetto Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-Stage Document Ranking with BERT. *CoRR* abs/1910.14424 (2019). arXiv:1910.14424 http://arxiv.org/abs/1910.14424

[39] Tiago Pimentel and Ryan Cotterell. 2021. A Bayesian Framework for Information-Theoretic Probing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 2869–2887. https://doi.org/10.18653/v1/2021.emnlp-main.229

[40] Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to Form: Measuring Systematicity as Information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1751–1764. https://doi.org/10.18653/v1/P19-1171

[41] Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 2481–2498. https://doi.org/10.18653/V1/2023.ACL-LONG.140

[42] Ori Ram, Liat Bezalel, Adi Zicher, Yonatan Belinkov, Jonathan Berant, and Amir Globerson. 2023. What Are You Token About? Dense Retrieval as Distributions Over the Vocabulary. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, 2481–2498. https://doi.org/10.18653/V1/2023.ACL-LONG.140

[43] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7237–7256. https://doi.org/10.18653/v1/2020.acl-main.647

[44] Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, Arianna Bisazza and Omri Abend (Eds.). Association for Computational Linguistics, 194–209. https://doi.org/10.18653/v1/2021.conll-1.15

[45] Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear Adversarial Concept Erasure. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 18400–18421. https://proceedings.mlr.press/v162/ravfogel22a.html

[46] Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. 2021. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance?. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (Eds.). Association for Computational Linguistics, 3363–3377. https://doi.org/10.18653/v1/2021.eacl-main.295

[47] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[48] Daniël Rennings, Lijun Lyu, and Avishek Anand. 2023. Listwise Explanations for Ranking Models using Multiple Explainers. In *Advances in Information Retrieval - 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, Proceedings, Part I (Lecture Notes in Computer Science)*. Springer.

[49] Daniël Rennings, Felipe Moraes, and Claudia Hauff. 2019. An Axiomatic Approach to Diagnosing Neural IR Models. In *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 11437)*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer, 489–503. https://doi.org/10.1007/978-3-030-15712-8_32

[50] Stephen E. Robertson and Karen Spärck Jones. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* 27, 3 (1976), 129–146. https://doi.org/10.1002/asi.4630270302

[51] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994 (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126. http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

[52] Julia Rozanova, Marco Valentino, Lucas C. Cordeiro, and André Freitas. 2023. Interventional Probing in High Dimensions: An NLI Case Study. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, 2444–2455. https://doi.org/10.18653/v1/2023.findings-eacl.188

[53] Mansi Sakarvadia, Arham Khan, Aswathy Ajith, Daniel Grzenda, Nathaniel Hudson, André Bauer, Kyle Chard, and Ian T. Foster. 2023. Attention Lens: A Tool for Mechanistically Interpreting the Attention Head Information Retrieval Mechanism. *CoRR* abs/2310.16270 (2023). https://doi.org/10.48550/ARXIV.2310.16270 arXiv:2310.16270

[54] Procheta Sen, Debasis Ganguly, Manisha Verma, and Gareth J. F. Jones. 2020. The Curious Case of IR Explainability: Explaining Document Scores within and across Ranking Models. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 2069–2072. https://doi.org/10.1145/3397271.3401286

[55] Jaspreet Singh and Avishek Anand. 2018. Posthoc interpretability of learning to rank models using secondary training data. *ArXiv preprint* abs/1806.11330 (2018). https://arxiv.org/abs/1806.11330

[56] Jaspreet Singh and Avishek Anand. 2020. Model agnostic interpretability of rankers via intent modelling. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna (Eds.). ACM, 618–628. https://doi.org/10.1145/3351095.3375234

[57] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html

[58] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. 2020. The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 107–118.

https://doi.org/10.18653/v1/2020.emnlp-demos.15

[59] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=SJzSgnRcKX

[60] Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. Function Vectors in Large Language Models. *CoRR* abs/2310.15213 (2023). https://doi.org/10.48550/ARXIV.2310.15213 arXiv:2310.15213

[61] Dan Vanderkam, Rob Schonberger, Henry Rowley, and Sanjiv Kumar. 2013. *Nearest Neighbor Search in Google Correlate*. Technical Report. Google. http://www.google.com/trends/correlate/nnsearch.pdf

[62] Elena Voita and Ivan Titov. 2020. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 183–196. https://doi.org/10.18653/v1/2020.emnlp-main.14

[63] Michael Völske, Alexander Bondarenko, Maik Fröbe, Benno Stein, Jaspreet Singh, Matthias Hagen, and Avishek Anand. 2021. Towards Axiomatic Explanations for Neural Ranking Models. In *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, Faegheh Hasibi, Yi Fang, and Akiko Aizawa (Eds.). ACM, 13–22. https://doi.org/10.1145/3471158.3472256

[64] Jonas Wallat, Fabian Beringer, Abhijit Anand, and Avishek Anand. 2023. Probing BERT for Ranking Abilities. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13981)*, Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo (Eds.). Springer, 255–273. https://doi.org/10.1007/978-3-031-28238-6_17

[65] Jonas Wallat, Jaspreet Singh, and Avishek Anand. 2020. BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online, 174–183. https://doi.org/10.18653/v1/2020.blackboxnlp-1.17

[66] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, Asli Celikyilmaz and Tsung-Hsien Wen (Eds.). Association for Computational Linguistics, 87–94. https://doi.org/10.18653/V1/2020.ACL-DEMOS.12

[67] Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards Explainable Search Results: A Listwise Explanation Generator. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 669–680. https://doi.org/10.1145/3477495.3532067

[68] 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 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 1941–1944. https://doi.org/10.1145/3397271.3401325

[69] Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and predict, and then predict again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 418–426.