Eclipse: Leveraging Pseudo-Irrelevance for Dimension Importance Estimation*

Giulio D'Erasmo[1,*], Giovanni Trappolini[1], Nicola Tonellotto[2] and Fabrizio Silvestri[1]

[1]Sapienza University of Rome, Rome, Italy
[2]University of Pisa, Pisa, Italy

Abstract

Recent advances in Information Retrieval (IR) have utilized high-dimensional embedding spaces to enhance the retrieval of relevant documents. The Manifold Clustering Hypothesis suggests that, although document embeddings are high-dimensional, the documents relevant to a specific query lie on a lower-dimensional manifold that depends on the query. This idea has motivated new retrieval methods, but current approaches still find it hard to clearly separate relevant signals from irrelevant noise. To address this issue, we present a new method called Eclipse, which uses information from both relevant and non-relevant documents. Our method calculates a centroid from the non-relevant documents and uses it as a reference to detect and estimate noisy dimensions in the relevant ones, leading to better retrieval results. Extensive experiments on three in-domain and one out-of-domain benchmarks demonstrate an average improvement of up to 21.03% (resp. 22.88%) in mAP(AP) and 12.04% (resp. 14.18%) in nDCG@10 w.r.t. the DIME-based baseline (resp. the baseline using all dimensions). Our results pave the way for more robust, pseudo-irrelevance-based retrieval systems in future IR research. We make the code available on Github[1].

Keywords

Dimension Importance Estimation, Relevance Feedback

1. Introduction

Dense retrieval models [17, 12, 18] embed queries and documents into a latent space with many dimensions, where vector similarities capture nuanced semantic relationships [19, 20]. However, while some dimensions encode meaningful semantic distinctions, others may introduce noise or contain non-discriminative information [7, 1, 4]. To address this issue, Dimension Importance Estimation (DIME) [14] was developed to identify and retain only the most informative dimensions, aiming to enhance retrieval performance by filtering out those that either contribute little or mostly capture noise [2, 8, 21]. Although DIME emphasizes relevant dimensions, the impact of irrelevant dimensions-those that add noise or non-discriminative information-remains largely unexplored. Existing methods, such as Rocchio's algorithm [26], show that improving a query involves adjusting it to be more centered around relevant documents, while also making it as far away as possible from irrelevant documents. We identify that explicitly modeling both relevant and irrelevant feedback can significantly improve dimension selection, thus improving dense retrieval performance. We introduce Eclipse, a novel method that utilizes representations of both relevant and irrelevant documents to more accurately identify important dimensions. In this paper, we explore how leveraging non-relevant documents through irrelevant feedback can improve state-of-the-art DIME approaches. We evaluate *ECLIPSE* across state-of-the-art TREC collections (Deep Learning 2019 [10], 2020 [9], DL-HARD 2021 [22], and Robust 2004 [28]), demonstrating improvements of up to 21.03% (resp. 22.88%) in mAP(AP) and 12.04% (resp. 14.18%) in nDCG@10 w.r.t the DIME-based baseline (resp. the baseline using all dimensions).

[1]https://github.com/giulio-derasmo/ECLIPSE-Contrastive-DIME-with-Pseudo-Irrelevance-Feedback

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*This is an extended abstract of the paper accepted at ICTIR 2025 [11].

*Corresponding author.

g.derasmo@diag.uniroma1.it (G. D'Erasmo); trappolini@diag.uniroma1.it (G. Trappolini); nicola.tonellotto@unipi.it (N. Tonellotto); fsilvestri@diag.uniroma1.it (F. Silvestri)

0009-0002-2287-051X (G. D'Erasmo); 0000-0002-5515-634X (G. Trappolini); 0000-0002-7427-1001 (N. Tonellotto); 0000-0001-7669-9055 (F. Silvestri)

2. Background and Preliminaries

In this section, we begin by outlining the classical Relevance Feedback model introduced by Rocchio [26], followed by a comprehensive overview of the Dimension Importance Estimation paradigm.

Rocchio. Rocchio's algorithm is a foundational method in information retrieval, refining query vectors by pulling them toward relevant documents and pushing them away from irrelevant ones. As modern IR systems rely on high-dimensional embeddings, moving beyond traditional vector space models requires exploring how to identify an optimal subset of query dimensions, rather than solely optimizing entire query vectors.

Dimension Importance Estimation (DIME). Faggioli et al. suggest that queries and documents exist in a lower-dimensional, query-dependent subspace of their high-dimensional latent space R^d. By projecting embeddings onto this subspace, a dense IR system can retain only the most informative dimensions for distinguishing relevance. DIMEs assign importance scores to dimensions using a query-dependent function. This score allows the system to rank the dimensions, retaining those with higher scores and discarding the less important ones. The selected dimensions thus form a low-dimensional, query-dependent subspace of $\mathbb{R}^d$. Two methods for estimating the importance of dimensions are PRF DIME and LLM DIME. The PRF DIME method utilizes pseudo-relevance feedback by assuming that the top-k documents retrieved by a similarity measure, such as BM25 [25], are likely relevant to the query [26, 30]. These documents are combined into a centroid vector $\mathbf{p}$ used to captures the alignment to the query $\mathbf{q}$, helping to rank and select the most relevant dimensions. LLM DIME, on the other hand, uses a synthetic document $\mathbf{a}$, generated by an LLM [12, 24, 16, 23, 3, 5, 27], assumed to be relevant to the query.

3. Our Method: Eclipse

In this section we introduce Eclipse, a novel framework designed to improve dense vector retrieval by including non-relevant documents in the decision-making of dimension importance estimation.

Formally, for a given query $\mathbf{q} \in \mathrm{R}^d$, which is embedded in a latent space using a bi-encoder, we follow the same procedure as in DIME to retrieve a set of k documents from the corpus. These documents are ranked using similarity measures such as cosine similarity or inner product. This set of documents, denoted as $\mathcal{D}_q = \{\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_k\}$, contains pseudo-relevant documents, whose content captures mainly relevant information and typically found at the top positions, and potentially pseudo-irrelevant documents at the bottom positions, whose content captures mainly irrelevant information. Now, fixing a parameter $0 < k^- < k$, we can define pseudo-irrelevant feedback by aggregating the embeddings of the bottom k^- documents in $\mathcal{D}_q$ into an irrelevant representative embedding $\overline{\mathbf{p}}$ as:

$$\overline{\mathbf{p}} = \frac{1}{k^-} \sum_{i=0}^{k^- - 1} \mathbf{d}_{k-i}.$$

We define Eclipse as a weighted difference between a pseudo-relevant representative embedding $\mathbf{p}^*$ and the irrelevant representative embedding $\overline{\mathbf{p}}$ as:

$$\overline{u}_q^*(i) = \alpha(\mathbf{q}_i \cdot \mathbf{p}_i^*) - \beta(\mathbf{q}_i \cdot \overline{\mathbf{p}}_i). \tag{1}$$

In Eq. (1), the embedding $\mathbf{p}^*$ depends on the original DIME used to compute the relevant signal. This formulation allows for the extension of any framework of DIME. Using pseudo-relevant feedback we can instantiate the vector $\mathbf{p}^{PRF}$ by aggregating the top $0 < k^+ < k - k^-$ document embeddings from $\mathcal{D}_q$ as: $\mathbf{p}^{PRF} = \frac{1}{k_+} \sum_{i=1}^{k^+} \mathbf{d}_i$.

We can also instantiate an LLM-based approach using the following pipeline: (1) Zero-shot prompting an LLM using the query q; (2) Use an encoder to embed the generated text into a latent vector representation $\mathbf{a} \in \mathrm{R}^d$; (3) Set $\mathbf{p}^{LLM} = \mathbf{a}$.

Table 1
Effectiveness metrics of our methods Eclipse ($\overline{u}^{PRF}$, $\overline{u}^{LLM}$) and baselines on different query sets and bi-encoders. In bold, the best performance observed for each triple IR system, test collection, and evaluation measure. Superscripts ^{a} and ^{b} indicate that the result is statistically significantly (p < 0.05) better than Baseline or standard DIMEs, respectively.

		AP					nDCG@10					AP					nDCG@10				
	Retained	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1	0.2	0.4	0.6	0.8	1
		DL '19										DL '20									
ANCE	u^{PRF}	.032	.253	.339	.370		.082	.552	.639	.657		.083	.287	.364	.390		.154	.541	.616	.651	
	u^{LLM}	.039	.267	.351	.370	.361	.104	.570	.660	.663	.645	.086	.281	.372	.397	.392	.174	.533	.622	.658	.646
	$\overline{u}^{PRF}$	.264^{b}	.361^{b}	.400^{ab}	.412^{ab}		.540^{b}	.624^{b}	.658^{b}	.674^{a}		.270^{b}	.364^{b}	.399^{b}	.413^{ab}		.477^{b}	.607^{b}	.627	.637	
	$\overline{u}^{LLM}$	.239^{b}	.380^{b}	.415^{ab}	**.435**^{ab}		.507^{b}	.684^{b}	**.719**^{ab}	.702^{ab}		.243^{b}	.346^{b}	.399^{ab}	**.417**^{b}		.350^{b}	.578	.633^{a}	**.669**	
Contriever	u^{PRF}	.497	.507	.511	.509		.675	.683	.692	.689		.484	.489	.497	.495		.713	.701	.704	.693	
	u^{LLM}	.522	.523	.521	.519	.494	.731	.736	.733	.745	.677	.491	.500	.504	.501	.478	.697	.695	.697	.689	.666
	$\overline{u}^{PRF}$	.523	.535^{ab}	.542^{ab}	.535^{ab}		.696	.710	.701	.704		.492	.509^{ab}	.509^{a}	.511^{ab}		**.716**^{a}	.706^{a}	.700^{a}	.696^{a}	
	$\overline{u}^{LLM}$	.534	.554^{ab}	**.558**^{ab}	.547^{ab}		.734^{a}	.747^{ab}	.749^{ab}	**.751**^{a}		.497^{a}	.510^{ab}	**.517**^{ab}	.511^{ab}		.686	.688	.704^{a}	.709^{ab}	
TAS-B	u^{PRF}	.496	.509	.512	.507		.725	.737	.738	.735		.473	.486	.489	.489		.698	.705	.712	.706	
	u^{LLM}	.509	.527	.520	.514	.476	.748	.762	.758	.757	.719	.469	.495	.494	.495	.475	.693	.697	.697	.705	.685
	$\overline{u}^{PRF}$	.523^{ab}	.547^{ab}	.548^{ab}	.542^{ab}		.734	.745^{a}	.735	.736		.472	.501^{ab}	.510^{ab}	.506^{ab}		.697	**.726**^{ab}	.720^{a}	.717^{a}	
	$\overline{u}^{LLM}$	.514	.544^{a}	**.557**^{ab}	.545^{ab}		.750	**.775**^{ab}	.771^{ab}	.769^{ab}		.470	.503^{a}	.510^{ab}	**.512**^{ab}		.676	.701	.702	.716^{a}	
		DL HD										RB '04									
ANCE	u^{PRF}	.019	.128	.176	.180		.057	.278	.340	.335		.015	.084	.135	.148		.055	.269	.357	.383	
	u^{LLM}	.014	.125	.172	.186	.180	.040	.253	.329	.346	.334	.015	.082	.137	.148	.146	.063	.257	.381	.392	.385
	$\overline{u}^{PRF}$	.105^{b}	.168^{b}	.183	.195^{ab}		.220^{b}	.324^{b}	.333	.349		.106^{b}	.147^{b}	.172^{ab}	**.179**^{ab}		.255^{b}	.362^{b}	.390^{b}	.405^{ab}	
	$\overline{u}^{LLM}$	.122^{b}	.172^{b}	.200^{b}	**.214**^{a}		.239^{b}	.318^{b}	.353	**.370**^{a}		.065^{b}	.124^{b}	.161^{ab}	.175^{ab}		.226^{b}	.356^{b}	.419^{ab}	**.439**^{ab}	
Contriever	u^{PRF}	.247	.255	.252	.251		.395	.393	.387	.384		.243	.254	.256	.257		.476	.489	.491	.492	
	u^{LLM}	.256	.259	.262	.261	.241	.377	.376	.390	.392	.375	.253	.263	.261	.259	.239	.516	.527	.519	.517	.480
	$\overline{u}^{PRF}$	.243	.260	**.277**^{ab}	.274^{ab}		.380	.402^{a}	**.410**^{ab}	.408^{ab}		.256^{ab}	.267^{ab}	**.269**^{ab}	.266^{ab}		.479	.492	.496^{a}	.501^{ab}	
	$\overline{u}^{LLM}$	.248	.271	.274^{a}	.267^{a}		.367	.390	.400	.392		.254^{a}	.264^{a}	.262^{a}	.261^{ab}		.516^{a}	.526^{a}	**.529**^{ab}	.520^{a}	
TAS-B	u^{PRF}	.235	.243	.250	.256		.366	.383	.388	.394		.199	.219	.224	.222		.418	.442	.448	.448	
	u^{LLM}	.257	.261	.265	.260	.238	.389	.402	.408	.395	.374	.177	.214	.218	.218	.197	.428	.466	.469	.467	.428
	$\overline{u}^{PRF}$	.261^{b}	.275^{ab}	.279^{ab}	.278^{ab}		.385	.408^{ab}	.421^{ab}	.410^{a}		.222^{ab}	.232^{ab}	**.233**^{ab}	.232^{ab}		.440^{b}	.446^{a}	.451^{a}	.458^{a}	
	$\overline{u}^{LLM}$	.267	.282^{ab}	**.285**^{ab}	.279^{ab}		.385	.410^{a}	.407	**.423**^{ab}		.189^{b}	.206^{a}	.216^{a}	.222^{a}		.433	.456^{a}	.467^{a}	**.475**^{a}	

Lastly, the parameters $\alpha, \beta \in \mathbb{R}$ control the balance between the relevant and irrelevant document signals. Rather than using a convex combination, we apply independent weighting to each term. This method provides greater flexibility and demonstrates superior performance in our experiments.

4. Experimental Setup

In our experiments, we compare our proposed Eclipse against the state-of-the-art DIMEs for dense IR systems. We experiment with three dense retrieval models: ANCE [29], Contriever [16], and TAS-B [15], all of which have been fine-tuned using the MS MARCO [6] passage dataset.

Datasets. We evaluate our methodology on three widely used benchmark collections for in-domain evaluation: TREC Deep Learning 2019 (DL '19) [10], TREC Deep Learning 2020 (DL '20) [9], and Deep Learning Hard (DL HD) [22]. To assess the robustness we further evaluate Eclipse on out-of-domain data based on the TREC Robust '04 (RB '04) collection [28]. We evaluate the systems using standard metrics such as mean Average Precision (AP) and nDCG@10.

Hyperparameters. We define four primary hyperparameters that influence different aspects of the model's decision-making process: k^+, k^-, α, and β. The parameter $k^+ \in \{1, \dots, 10\}$ (resp. $k^- \in \{1, \dots, 14\}$) determines the number of relevant (resp. irrelevant) documents, used to build our pseudo-relevance embeddings. The hyperparameter α controls the strength of the relevant representative embedding, while β modulates the denoising effect of the irrelevant representative embedding. Both are positive values increasing linearly from 0.1 up to 1. For combinations where $\alpha = \beta$ we test the base case of $\alpha = \beta = 1$.

Baselines. We compare our method to standard DIMEs, PRF DIME and LLM DIME. We use GPT4 [13] as LLM in our experiments. We will refer to the dense IR system at full dimensionality as *Baseline*. All the DIMEs, including Eclipse version, use a retrieved collection of documents $\mathcal{D}_q$ of size $1,000$.

Table 2
Percentage improvement in Recall from LLM DIME to LLM Eclipse at different relevance levels for various bi-encoders across multiple datasets. Low relevance = 1, Medium relevance = 2, High relevance = 3.

	DL '19			DL '20			DL HD			RB '04	
Relevance	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	High
ANCE	16.91	12.86	8.74	2.43	2.34	2.10	15.65	0.09	1.35	12.78	6.53
Contriever	4.72	1.56	2.80	-1.49	-0.24	-0.20	4.53	0.75	-1.14	0.33	1.18
TAS-B	3.16	1.14	1.75	1.52	0.41	0.09	0.82	1.57	-0.85	0.39	1.33

5. Experiments

In our experiments, we investigate the following research questions: **RQ1**: Can non-relevant documents be leveraged using irrelevant feedback to improve state-of-the-art DIME approaches? **RQ2**: Are metrics of the retrieval pipeline impacted differently by nonrelevant results when used for dimension importance estimation?

Results for RQ1: Table 1 compare both versions of Eclipse with standard DIMEs (PRF and LLM) on the TREC DL '19, DL '20, DH, and RB '04 datasets, using the ANCE, Contriever, and TAS-B models. We report the performance using the best configuration for all the DIMEs (standards and Eclipse) in the table. The most interesting results is over ANCE, where Eclipsereduce the percentage of retained dimensions needed to surpass the baseline when using all the dimensions to just 40-60%, demonstrating that explicitly modeling both positive and negative feedback in the DIME framework yields a robust improvement. The gains are especially notable, with improvements of 21.03% in AP and 12.04% in nDCG@10 relative to DIMEs, and even higher margins over the standard baseline: 22.88% (AP) and 14.18% (nDCG@10).

Eclipse exhibits superior performance in the traditional evaluation protocol, improving performance up to 21.03% (resp. 22.88%) in AP and 12.04% (resp. 14.18%) in nDCG@10 w.r.t. the DIME-based baseline (resp. the baseline using all dimensions). In particular, both PRF Eclipse and LLM Eclipse show statistically significant improvement with respect to their DIME counterparts and Baseline.

Results for RQ2: To understand how the presence of nonrelevant documents in the dimension importance estimation pipeline affects different aspects of the retrieval pipeline, we analyzed the recall performance of LLM Eclipse compared the standard LLM DIME. Table 2 demonstrates that LLM Eclipse achieves consistent recall improvements over LLM DIME across multiple datasets and bi-encoders, with the most notable gains observed for low and medium relevance documents. This effect is especially pronounced in the DL collections, where recall increases of up to 16.91% are observed for marginally relevant documents. As a result, this explain why LLM Eclipse yields a larger boost in AP, which is sensitive to recall across all relevance levels. In contrast, improvements in nDCG@10 are more modest, reflecting the smaller gains for highly relevant documents that dominate the top-ranked results.

6. Conclusion and Future Work

We present Eclipse, a novel method designed to enhance dense retrieval by exploiting pseudo-irrelevant feedback. This approach offers improved separation between relevant and non-relevant dimensions within document embeddings. Unlike conventional DIME methods that rely solely on relevance signals, Eclipse introduces a contrastive perspective by utilizing irrelevant documents.

Eclipse achieves an average improvement of up to 21.03% (and 22.88% for AP) and 12.04% (and 14.18% for nDCG@10) compared to the DIME-based baseline (and the baseline using all dimensions).

By emphasizing relevant embedding dimensions, Eclipse promotes moderately relevant documents within the ranking, leading to marked gains in AP. Future research should focus on predicting a unique percentage of retained dimensions for each queries. Another unexplored section is the use of irrelevant documents generated by LLMs as a substitute for human-generated documents.

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] Khetam Al Sharou, Zhenhao Li, and Lucia Specia. Towards a better understanding of noise in natural language processing. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online, September 2021. INCOMA Ltd.

[2] Sileye 0. Ba. Discovering topics with neural topic models built from plsa assumptions, 2019.

[3] Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri. Dantellm: Let's push italian llm research forward! In *Proceedings of the 2024 Joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 4343–4355, 2024.

[4] Andrea Bacciu, Florin Cuconasu, Federico Siciliano, Fabrizio Silvestri, Nicola Tonellotto, and Giovanni Trappolini. Rraml: Reinforced retrieval augmented machine learning. volume 3537, page 29 – 37, 2023. Cited by: 7.

[5] Andrea Bacciu, Giovanni Trappolini, Andrea Santilli, Emanuele Rodolà, and Fabrizio Silvestri. Fauno: The italian large language model that will leave you senza parole! volume 3448, page 9 – 17, 2023. Cited by: 7.

[6] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, and Xiaodong Liu et al. Ms marco: A human generated machine reading comprehension dataset, 2018.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[8] Happy Buzaaba and Toshiyuki Amagasa. A scheme for efficient question answering with low dimension reconstructed embeddings. In *The 23rd International Conference on Information Integration and Web Intelligence*, iiWAS2021, page 303–310, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the trec 2020 deep learning track, 2021.

[10] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2019 deep learning track, 2020.

[11] Giulio D'Erasmo, Giovanni Trappolini, Fabrizio Silvestri, and Nicola Tonellotto. Eclipse: Contrastive dimension importance estimation with pseudo-irrelevance feedback for dense retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 147–154, New York, NY, USA, 2025. Association for Computing Machinery.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Josh Achiam et al. Gpt-4 technical report, 2024.

[14] Guglielmo Faggioli, Nicola Ferro, Raffaele Perego, and Nicola Tonellotto. Dimension importance estimation for dense information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1318–1328, New York, NY, USA, 2024. Association for Computing Machinery.

[15] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling, 2021.

[16] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[18] Omar Khattab and Matei Zaharia. 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 '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery.

[19] Wen Li, Ying Zhang, Yifang Sun, Wei Wang, Mingjie Li, Wenjie Zhang, and Xuemin Lin. Approximate nearest neighbor search on high dimensional data — experiments, analyses, and improvement. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1475–1488, 2020.

[20] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 04 2021.

[21] Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. Simple and effective unsupervised redundancy elimination to compress dense vectors for passage retrieval. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2854–2859, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[22] Iain Mackie, Jeffrey Dalton, and Andrew Yates. How deep is your learning: the dl-hard annotated deep learning dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2335–2341, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, 2018.

[24] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[25] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, April 2009.

[26] J.J. Rocchio. *Relevance Feedback in Information Retrieval.* Prentice Hall, Englewood Cliffs, New Jersey, 1971.

[27] Gabriele Tolomei, Cesare Campagnano, Fabrizio Silvestri, and Giovanni Trappolini. Prompt-to-os (p2os): revolutionizing operating systems and human-computer interaction with integrated ai generative models. In *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 128–134. IEEE, 2023.

[28] Ellen M. Voorhees. Overview of the trec 2004 robust track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, Gaithersburg, MD, 2004. NIST Special Publication 500-261, National Institute of Standards and Technology (NIST).

[29] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021.

[30] Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Trans. Inf. Syst.*, 18(1):79–112, January 2000.