Correlation Between Attention Map Sparsity and Retrieval Recall in Domain-Adapted Dense Retrievers on ScienceQA

Assignee Research

June 11, 2026

Abstract

Dense retrievers have demonstrated significant potential for neural information retrieval; however, they exhibit a lack of robustness to domain shifts, thereby limiting their efficacy in zero-shot settings across diverse domains. Previous research has investigated unsupervised domain adaptation techniques to adapt dense retrievers to target domains. However, these studies have not focused on explainability analysis to understand how such adaptations alter the model's behavior. In this paper, we propose utilizing the integrated gradients framework to develop an interpretability method that prov

1 Introduction

This paper examines: Interpretability Analysis of Domain Adapted Dense Retrievers. Research question: What is the correlation between attention map sparsity in domain-adapted dense retrievers and their retrieval recall metrics on the ScienceQA multimodal benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.3/10.

3 Results

15 papers retrieved. 12 claims extracted; 11 independently verified. Quality review score: 8.3/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
Domain adaptation improves NDCG@10 from 65.1 to 71.6 (+6.5 abs., +10%) on TREC-COVID.	✓	0.20
Domain adaptation improves NDCG@10 from 26.7 to 36.8 (+10.1 abs., +38%) on FIQA.	✓	0.19
The DistilBERT model effectively matches query terms with document terms, assigning high positive attribution scores to	✓	0.27
The model assigns no attributions for the [CLS] and [SEP] tokens.	✓	0.15
The model identifies 'corona' and 'disease' as important words despite the query not explicitly mentioning them.	✓	0.15
The term 'complications' appears in the negative attribution word cloud, which is a term present in the query text.	✓	0.19
For the FIQA dataset, the model assigns positive attributions to document tokens ['what', 'is', 'french'] and negative a	✓	0.25
The term 'gold' appears among the negatively contributing terms for the FIQA dataset.	✓	0.18
After domain adaptation, the query contribution of 'complications' has decreased compared to the non-domain-adapted mode	✓	0.25
The domain-adapted model assigns more importance to the first sentence, which is the title.	×	0.13
Domain-adapted models focus more on in-domain terminology compared to non-adapted models, exemplified by terms such as '	✓	0.32
Integrated gradients are a viable choice for explaining and analyzing the internal mechanisms of dense retrievers.	✓	0.28

References

- http://arxiv.org/abs/2507.23334v2
- http://arxiv.org/abs/2501.14459v1
- http://arxiv.org/abs/2404.14464v1