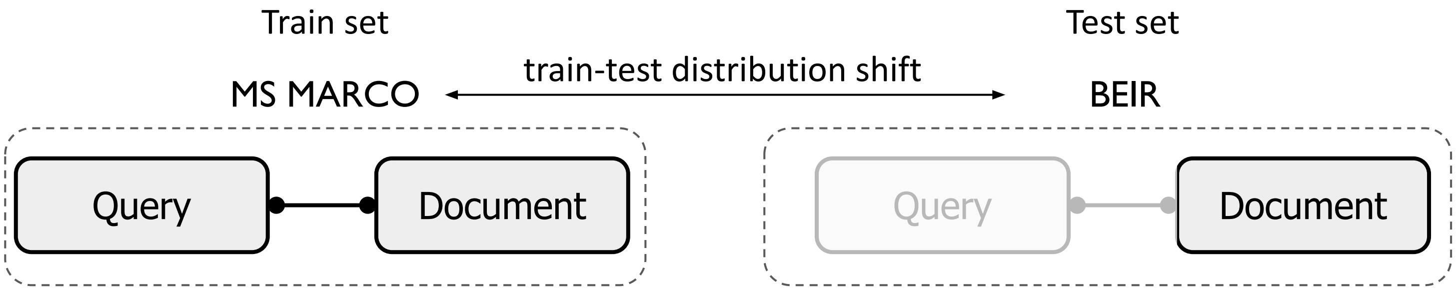
Relevance-assisted Generation for Robust Zero-shot Retrieval

Jihyuk Kim, Minsoo Kim , Joonsuk Park, Seung-won Hwang
(**presenter: Jihyuk Kim**)

Task: zero-shot retrieval



✅ Relevance annotations are available for large-scale open-domain corpus

❌ Queries and relevance annotations are often not available, for specialized domains (e.g., medical question) → *zero-shot retrieval*

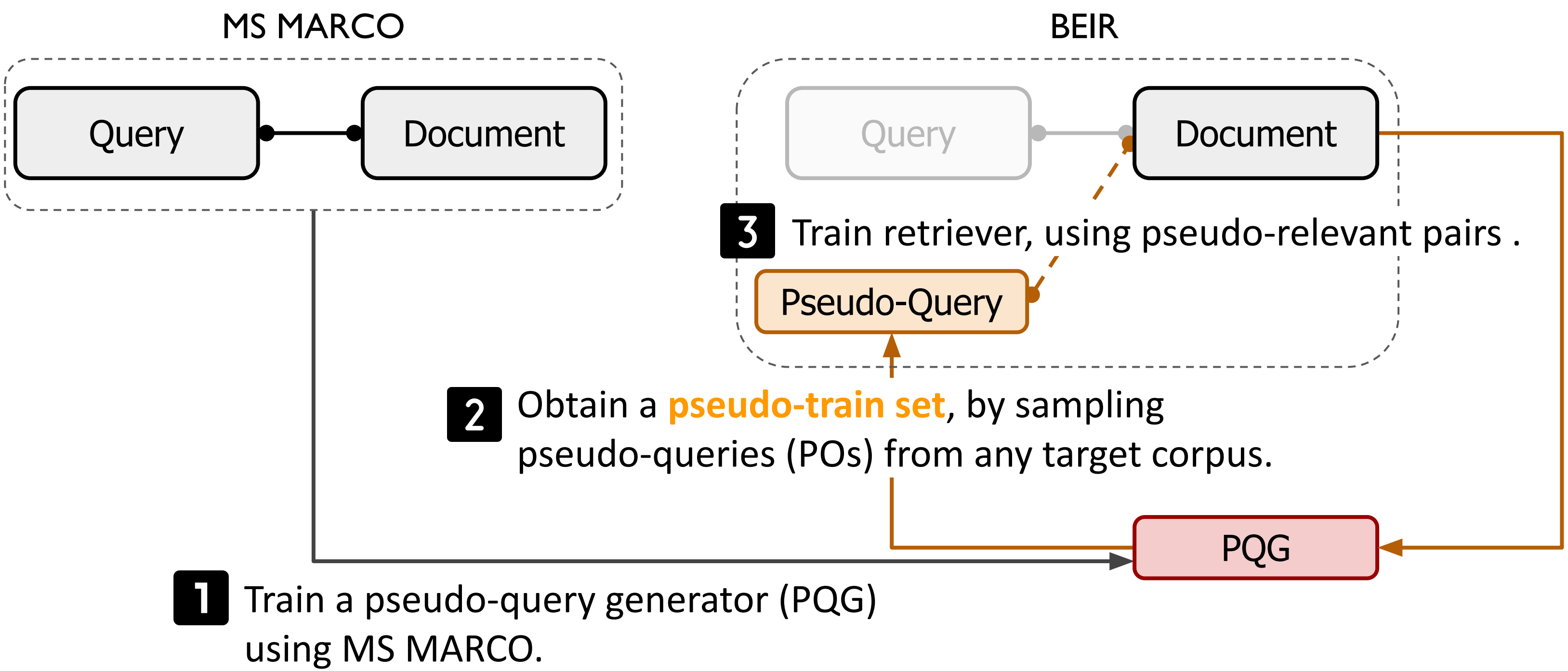
Dense retrievers do not generalize well to out-of-domain corpus.

🤔 RQ. How can we make the dense retriever domain-adaptive?

Create domain-tailored train set, via pseudo-query generation

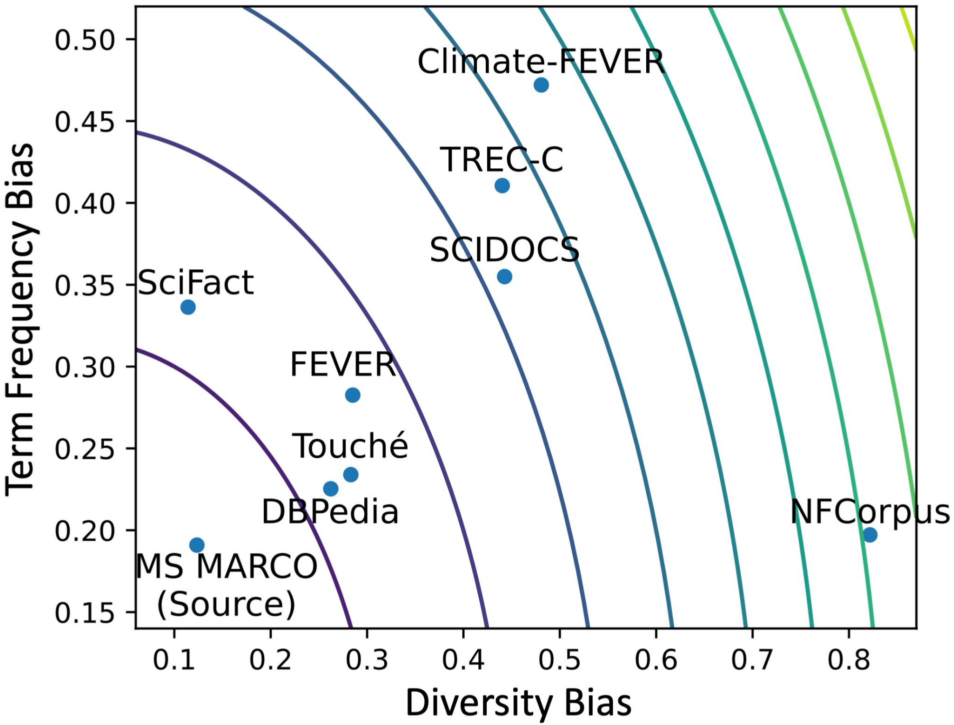
What if we can get a training dataset for any given corpus?

We can! via pseudo-query generation!



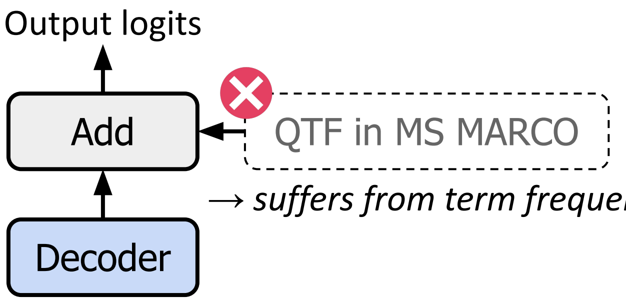
Challenge: distribution shift

- **Term frequency bias**:
  PQG fails to generate query terms, when rarely observed in MS MARCO.
- **Diversity bias**:
  A fixed-size single dense vector cannot cover diverse topics in long documents.



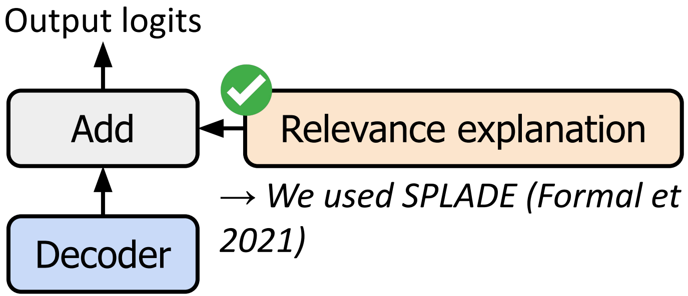
Proposed solution 1. relevance-assisted PQG

[Baseline PQG]

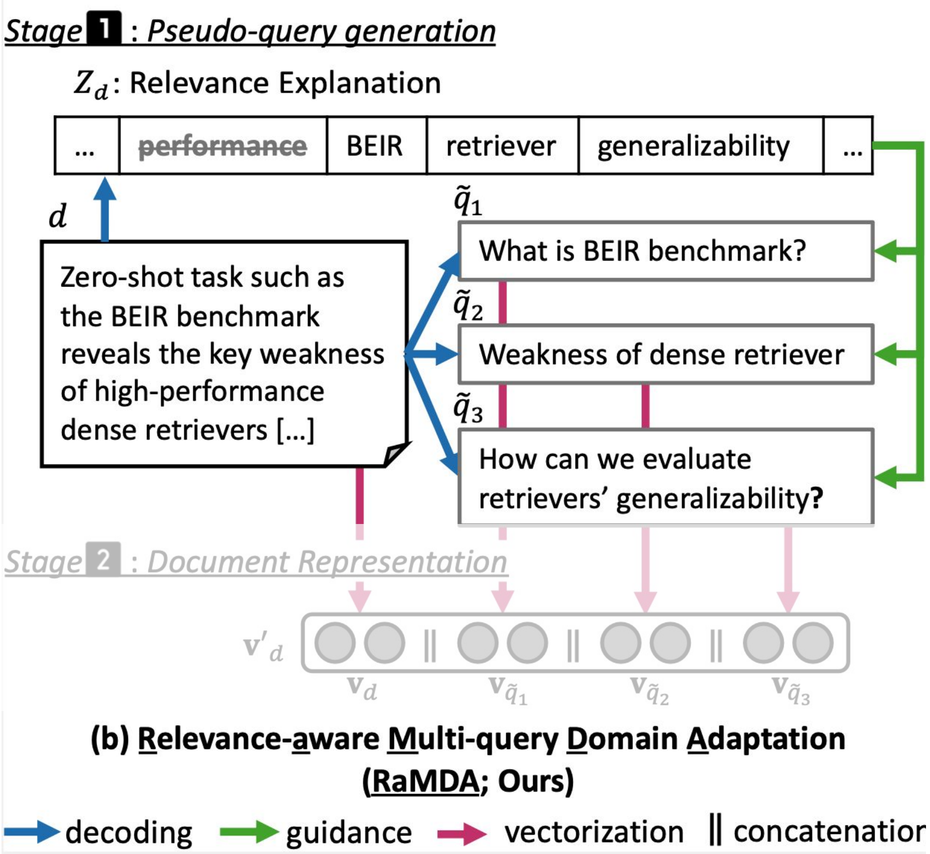


→ *suffers from term frequency bias*

[Relevance-assisted PQG]



→ *We used SPLADE (Formal et al., 2021)*



(b) Relevance-aware Multi-query Domain Adaptation (RaMDA; Ours)

Formal et al., 2021. "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking"

Proposed solution 2. multi-query vector augmentation

Hypothetical long document

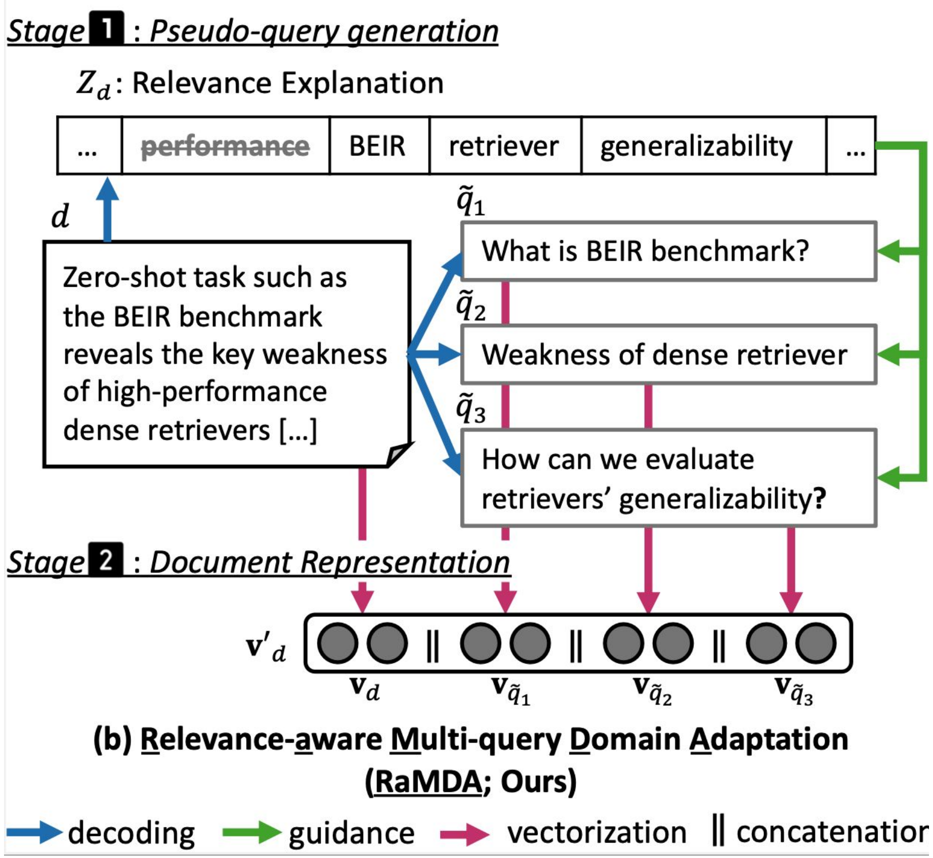
… BEIR benchmark reveals the key weakness …

dense retrievers poorly generalize to out-of-domain corpus …

…

Out-of-domain datasets have been released, enabling evaluation generalizability …

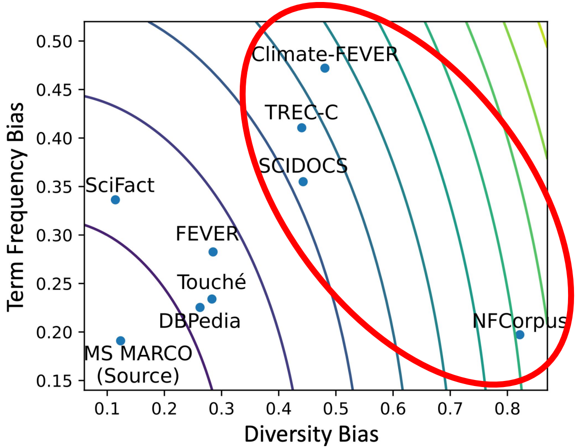
- *What is BEIR benchmark?*
- *What are the weaknesses of dense retrievers?*
- *How can we evaluate the generalization ability?*



(b) Relevance-aware Multi-query Domain Adaptation (RaMDA; Ours)

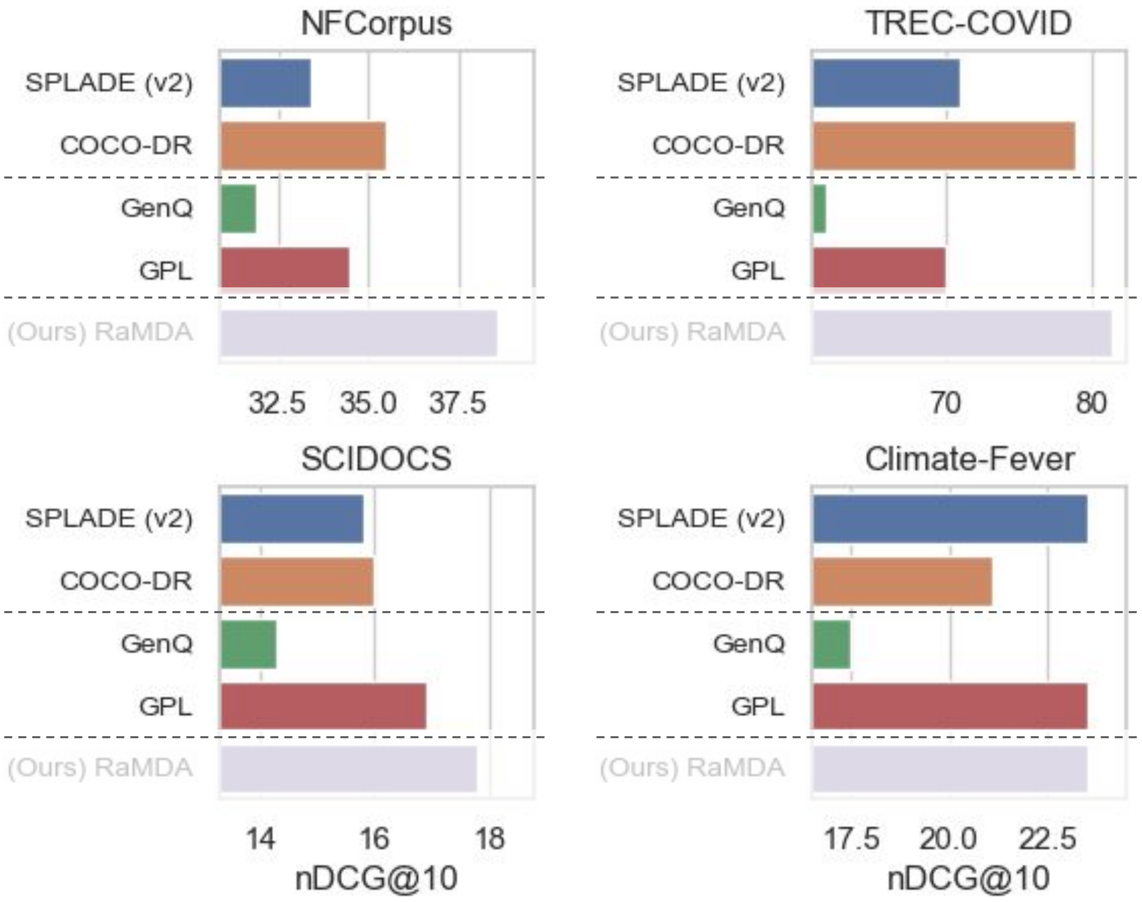
Experiment

- Dataset: Four BEIR datasets, showing the largest distribution shifts from MS MARCO
  - NFCorpus
  - SCIDOCS
  - TREC-COVID
  - Climate-FEVER



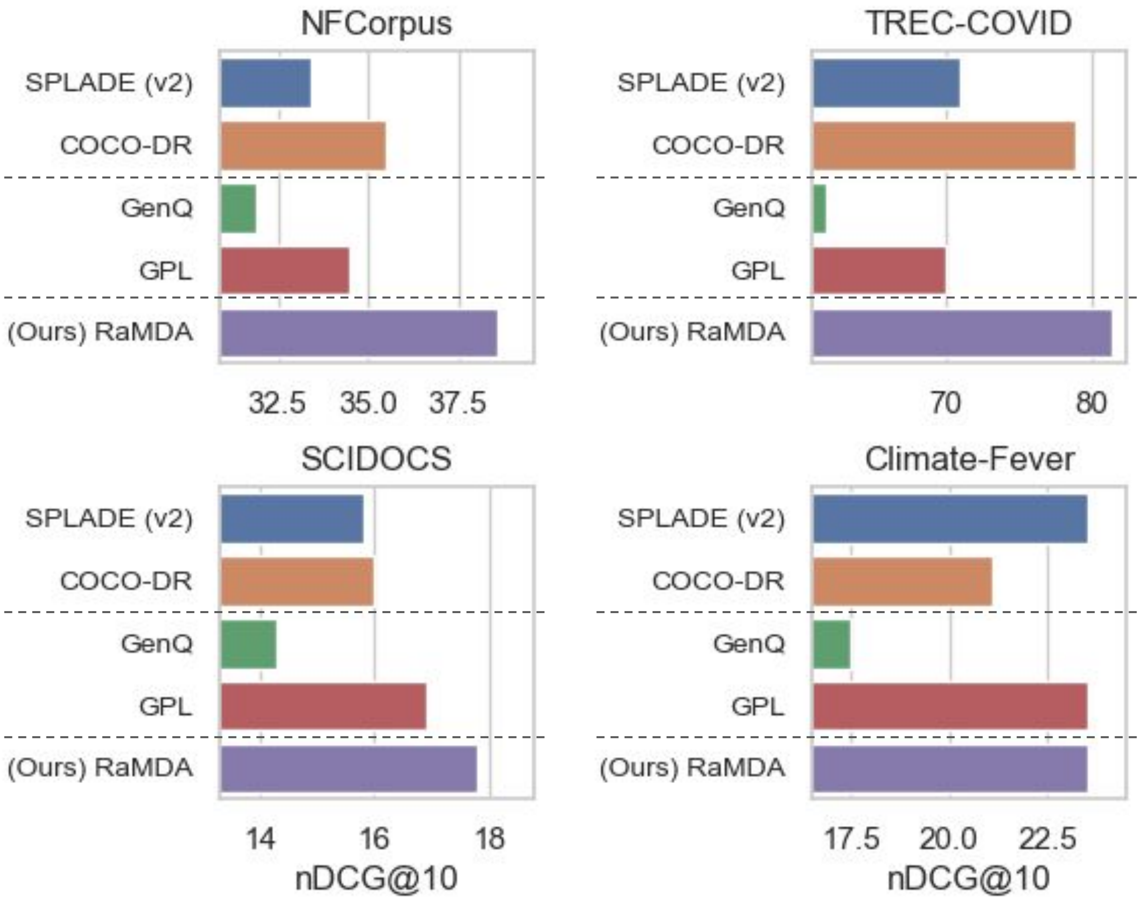
- Baseline
  - **GenQ** generates pseudo-queries from BEIR corpus, and fine-tunes a dense retriever.
  - **GPL**, building upon GenQ, additionally utilizes an expensive cross-encoder to label relevance of pseudo-queries.
  - *+ domain invariant retriever*: **COCO-DR** and **SPLADE**, as state-of-the-art dense and sparse retriever, respectively.

Results



Previous PQG approaches often underperform domain-invariant retrievers.

Results



Our model shows better or comparable performance than all baselines.

Thank you!