How Train-Test Leakage Affects Zero-shot Retrieval

Maik Fröbe\textsuperscript{1}  
Christopher Akiki\textsuperscript{2}  
Martin Potthast\textsuperscript{2}  
Matthias Hagen\textsuperscript{1}

Friedrich Schiller University Jena\textsuperscript{1}  
Leipzig University\textsuperscript{2}

SPIRE, 8–10 November 2022

webis.de
How Train-Test Leakage Affects Zero-shot Retrieval

Motivation: Leaderboard for Retrieval Effectiveness on Robust04

- Robust04: 249 test queries with dense judgments
  - Traditional setup with cross-validation
How Train-Test Leakage Affects Zero-shot Retrieval

Motivation: Leaderboard for Retrieval Effectiveness on Robust04

- Robust04: 249 test queries with dense judgments
  - Traditional setup with cross-validation
- MonoT5 (zero-shot)
  - Trained only on MS MARCO (> 10 million queries available)
  - There might be overlapping queries: Is this train–test leakage?
How Train-Test Leakage Affects Zero-shot Retrieval

Overlapping Queries for Topic 441 of Robust04

- Train on many queries
- Test on 249 queries

**Title:** lyme disease

**Description:** How do you prevent and treat Lyme disease?

**Narrative:** Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.

**Query variants:**
- lyme disease treatments
- prevent lyme disease
- ...
How Train-Test Leakage Affects Zero-shot Retrieval

Overlapping Queries for Topic 441 of Robust04

Train on many queries

Test on 249 queries
How Train-Test Leakage Affects Zero-shot Retrieval

Overlapping Queries for Topic 441 of Robust04

- Train on many queries
- Test on 249 queries

MS MARCO

- Title: lyme disease
- Description: How do you prevent and treat Lyme disease?
- Narrative: Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.

Robust04

- Query variants:
  - lyme disease treatments
  - prevent lyme disease
  ...
How Train-Test Leakage Affects Zero-shot Retrieval

Overlapping Queries for Topic 441 of Robust04

- Train on many queries
- Test on 249 queries

MS MARCO

Robust04

- Title: lyme disease
- Description: How do you prevent and treat Lyme disease?
- Narrative: Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.
- Query variants: lyme disease treatments prevent lyme disease ...

...
How Train-Test Leakage Affects Zero-shot Retrieval

Overlapping Queries for Topic 441 of Robust04

MS MARCO

- lyme disease
- how to treat lyme disease
- how to prevent lyme disease
- lyme disease treatment
- prevent lyme disease

Robust04

- Title: lyme disease
- Description: How do you prevent and treat Lyme disease?
- Narrative: Documents that discuss current prevention and treatment techniques for Lyme disease are relevant. Reports of research on new treatments of the disease are also relevant.
- Query variants: lyme disease treatments prevent lyme disease ...

- Train on many queries
- Test on 249 queries

Is the evaluation of MonoT5 invalidated by overlapping queries?
How Train-Test Leakage Affects Zero-shot Retrieval

Might MonoT5 Benefit From Overlapping Queries?

MonoT5

- 3 billion parameters sequence-to-sequence model
- The query $q$ and the document $d$ are embedded in a input sequence:

  Query: $q$  Document: $d$  Relevant:

- Documents ranked by the probability that the next token is “true”
Monot5

- 3 billion parameters sequence-to-sequence model
- The query $q$ and the document $d$ are embedded in an input sequence:

```
Query: q  Document: d  Relevant:
```

- Documents ranked by the probability that the next token is “true”
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss

Pilot Study

- We review 100 query-topic pairs to identify a precision-oriented threshold
- Candidates for overlapping queries:
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss

Pilot Study

- We review 100 query-topic pairs to identify a precision-oriented threshold
- Candidates for overlapping queries:

| Candidates    | Robust04 |
|---------------|----------|
|                | Topics   | Queries |
| Title          | 140      | 1,775    |
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss

Pilot Study

- We review 100 query-topic pairs to identify a precision-oriented threshold
- Candidates for overlapping queries:

| Candidates   | Robust04 |
|--------------|----------|
|              | Topics   | Queries |
| Title        | 140      | 1,775   |
| Description  | 8        | 50      |
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss

Pilot Study

- We review 100 query-topic pairs to identify a precision-oriented threshold
- Candidates for overlapping queries:

| Candidates       | Robust04 |
|------------------|----------|
|                  | Topics   | Queries |
| Title            | 140      | 1,775   |
| Description      | 8        | 50      |
| Variants         | 167      | 3,356   |
How Train-Test Leakage Affects Zero-shot Retrieval

Candidates for Leaking Queries

- Nearest-neighbor search for overlapping queries
- Sentence-BERT embeddings for all MS MARCO and ORCAS queries
- Exact cosine similarity nearest-neighbor search with Faiss

Pilot Study

- We review 100 query-topic pairs to identify a precision-oriented threshold
- Candidates for overlapping queries:

| Candidates | Robust04 |
|------------|----------|
|            | Topics   | Queries   |
| Title      | 140      | 1,775     |
| Description| 8        | 50        |
| Variants   | 167      | 3,356     |
| Union      | 181      | 3,960     |
How Train-Test Leakage Affects Zero-shot Retrieval

Verification of Candidates for Leaking Queries

- Manually review of the 5 most similar candidates per topic above threshold
- Identified query reformulation types:

| Type            | Queries |
|-----------------|---------|
| Identical       | 187     |
| Generalization  | 124     |
| Specialization  | 228     |
| Reformulation   | 182     |
| Different Topic | 106     |
How Train-Test Leakage Affects Zero-shot Retrieval

Verification of Candidates for Leaking Queries

- Manually review of the 5 most similar candidates per topic above threshold
- Identified query reformulation types:

| Type            | Queries |
|-----------------|---------|
| Identical       | 187     |
| Generalization  | 124     |
| Specialization  | 228     |
| Reformulation   | 182     |
| Different Topic | 106     |

172 of 249 test queries from Robust04 occur in MS MARCO (69%)
How Train-Test Leakage Affects Zero-shot Retrieval

Impact of Leaking Queries: Experimental Setup

- Models trained on dedicated datasets to assess train–test leakage
- Varying training set sizes: 1,000 to 128,000 instances
- Each model trained five times on each dataset

Training Datasets

- No Leakage
  - Random non-leaking queries
  - balanced between MS MARCO and ORCAS
- MS MARCO Leakage
  - 500 random manually verified leaking queries from MS MARCO
  - supplemented by no-leakage queries
- Test Leakage
  - 500 queries from the actual test data
  - supplemented by no-leakage queries
  - Meant as an “upper bound” for any train–test leakage effect
How Train-Test Leakage Affects Zero-shot Retrieval Effectiveness of Retrieval Models
How Train-Test Leakage Affects Zero-shot Retrieval Effectiveness of Retrieval Models

- Multiple models in five-fold cross-validation setup

| Model      | nDCG@10 on R04                  |       |       |
|------------|--------------------------------|-------|-------|
|            | No Leakage | MS MARCO Leakage | Test Leakage |
| Duet       | 0.201       | 0.198            | 0.224†       |
| KNRM       | 0.194       | 0.214†           | 0.309†       |
| monoBERT   | 0.394       | 0.373†           | 0.396        |
| monoT5     | 0.461       | 0.457            | 0.478†       |
| PACRR      | 0.382       | 0.364†           | 0.391        |
## How Train-Test Leakage Affects Zero-shot Retrieval Effectiveness of Retrieval Models

Increase in rank-offset between leaked relevant and non-relevant documents

| Model    | MS MARCO Leakage | Test Leakage |
|----------|------------------|--------------|
| Duet     | 6.378 ±32.15     | 0.809 ±17.69 |
| KNRM     | 0.640 ±19.22     | 1.335 ±11.75 |
| monoBERT | 0.692 ±17.97     | 3.886 ±20.39 |
| monoT5   | 0.443 ±8.60      | 3.443 ±19.96 |
| PACRR    | 0.043 ±19.30     | 1.952 ±17.71 |
How Train-Test Leakage Affects Zero-shot Retrieval

Takeaways

- Possible train–test leakage for models trained on MS MARCO
  - Potential to invalidate experiments
  - Default in PyTerrier/Pyserini/PyGaggle often trained on MS MARCO
  - Only few training instances overlap: Impact measurable, but negligible

- Future work:
  - Effects on Dense Retrieval models
  - Practical consequences for real search engines
How Train-Test Leakage Affects Zero-shot Retrieval

Takeaways

- Possible train–test leakage for models trained on MS MARCO
  - Potential to invalidate experiments
  - Default in PyTerrier/Pyserini/PyGaggle often trained on MS MARCO
  - Only few training instances overlap: Impact measurable, but negligible

- Future work:
  - Effects on Dense Retrieval models
  - Practical consequences for real search engines

Thank You!