# Revisiting Query Variation Robustness of Transformer Models

November 12 – 16, 2024

**Tim Hagen** Harry Scells Martin Potthast

University of Kassel and hessian.AI

UNIKASSEL VERSITÄT

hessian.AI

# The Problem

- $\approx 70\,\%$ of information seeking queries are **keyword queries**
- $26\,\%$ of queries contain **typos**

# The Problem

- $\approx 70\,\%$ of information seeking queries are **keyword queries**
- $26\,\%$ of queries contain **typos**
- Transformer-based rankers have been shown not to be robust to using keywords and typos

# The Problem

- $\approx 70\,\%$ of information seeking queries are **keyword queries**
- $26\,\%$ of queries contain **typos**
- Transformer-based rankers have been shown not to be robust to using keywords and typos
- Previous work mostly focussed on typos

# The Problem

- $\approx 70\,\%$ of information seeking queries are **keyword queries**
- $26\,\%$ of queries contain **typos**
- Transformer-based rankers have been shown not to be robust to using keywords and typos
- Previous work mostly focussed on typos

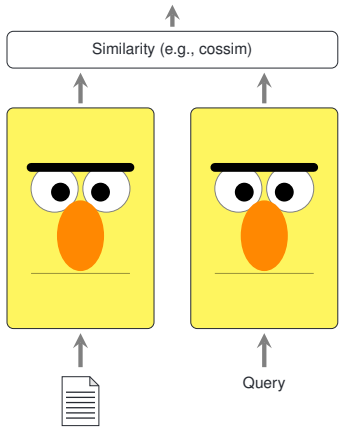
**Research Question**

How robust are more recent transformer-based language models?

# Background

## Dense Retrieval

- Using embedding models for ranking:
  - Embed query and document separately
  - Rank using the cossim of the documents' embeddings to the query's
- Transformer-based ranking models are the first neural architecture to demonstrably outperform traditional approaches



# Experiments

## Models

**SBERT**

- Popular embedding model
- Based on DistilBERT$_{\text{Base}}$
- 66M parameters

# Experiments

## Models

**SBERT**

- Popular embedding model
- Based on DistilBERT$_{\text{Base}}$
- 66M parameters

**CharacterBERT-DR-ST**

- Typo-aware architecture & pre-training
- Based on BERT$_{\text{Base}}$
- 104M parameters

# Experiments

## Models

**SBERT**

- Popular embedding model
- Based on DistilBERT$_{\text{Base}}$
- 66M parameters

**CharacterBERT-DR-ST**

- Typo-aware architecture & pre-training
- Based on BERT$_{\text{Base}}$
- 104M parameters

**E5 Mistral**

- #1 on MTEB[1]
- Based on Mistral-7B-instruct
- 7B parameters

[1] At the time of our experiments

# Experiments

## Models

### SBERT

- Popular embedding model
- Based on DistilBERT$_{\text{Base}}$
- 66M parameters

### CharacterBERT-DR-ST

- Typo-aware architecture & pre-training
- Based on BERT$_{\text{Base}}$
- 104M parameters

### E5 Mistral

- #1 on MTEB[1]
- Based on Mistral-7B-instruct
- 7B parameters

### AnglE

- #2 on MTEB[1]
- Based on BERT$_{\text{Large}}$
- 335M parameters

[1] At the time of our experiments

# Experiments

## Models

### SBERT

- Popular embedding model
- Based on DistilBERT$_{\text{Base}}$
- 66M parameters

### CharacterBERT-DR-ST

- Typo-aware architecture & pre-training
- Based on BERT$_{\text{Base}}$
- 104M parameters

### E5 Mistral

- #1 on MTEB[1]
- Based on Mistral-7B-instruct
- 7B parameters

### AnglE

- #2 on MTEB[1]
- Based on BERT$_{\text{Large}}$
- 335M parameters

### Ada v2

- SOTA commercial embedding model

*No architectural details publicly available*

[1] At the time of our experiments

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

| Query variation | | Example | # Queries | |
|---|---|---|---|---|
| Category | Transform. heuristic | | TREC DL '19 | ANTIQUE |
| Original | | what is durable medical equipment consist of | 43 | 200 |

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

| Query variation | | Example | # Queries | |
|---|---|---|---|---|
| Category | Transform. heuristic | | TREC DL '19 | ANTIQUE |
| Original | | what is durable medical equipment consist of | 43 | 200 |
| | NeighbCharSwap | what is durable **mdeical** equipment consist of | 43 | 199 |
| Misspelling | RandomCharSub | what is durable **medycal** equipment consist of | 42 | 197 |
| | QWERTYCharSub | what is durable medical equipment **xonsist** of | 42 | 182 |

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

| Query variation | | Example | # Queries | |
|---|---|---|---|---|
| Category | Transform. heuristic | | TREC DL '19 | ANTIQUE |
| Original | | what is durable medical equipment consist of | 43 | 200 |
| Misspelling | NeighbCharSwap | what is durable **mdeical** equipment consist of | 43 | 199 |
| | RandomCharSub | what is durable **medycal** equipment consist of | 42 | 197 |
| | QWERTYCharSub | what is durable medical equipment **xonsist** of | 42 | 182 |
| Naturality | RemoveStopWords | ~~**what is**~~ durable medical equipment consist ~~**of**~~ | 37 | 199 |
| | T5DescToTitle | ~~**what is**~~ durable medical equipment ~~**consist of**~~ | 35 | 136 |

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

| Query variation | | Example | # Queries | |
|---|---|---|---|---|
| Category | Transform. heuristic | | TREC DL '19 | ANTIQUE |
| Original | | what is durable medical equipment consist of | 43 | 200 |
| Misspelling | NeighbCharSwap | what is durable **mdeical** equipment consist of | 43 | 199 |
| | RandomCharSub | what is durable **medycal** equipment consist of | 42 | 197 |
| | QWERTYCharSub | what is durable medical equipment **xonsist** of | 42 | 182 |
| Naturality | RemoveStopWords | ~~**what is**~~ durable medical equipment consist ~~**of**~~ | 37 | 199 |
| | T5DescToTitle | ~~**what is**~~ durable medical equipment ~~**consist of**~~ | 35 | 136 |
| Ordering | RandomOrderSwap | **medical** is durable **what** equipment consist of | 43 | 200 |

# Experiments

## Dataset

- Query variation dataset by Penha et al.
- **Semantically equivalent** query variations

| Query variation | | Example | # Queries | |
|---|---|---|---|---|
| Category | Transform. heuristic | | TREC DL '19 | ANTIQUE |
| Original | | what is durable medical equipment consist of | 43 | 200 |
| Misspelling | NeighbCharSwap | what is durable **mdeical** equipment consist of | 43 | 199 |
| | RandomCharSub | what is durable **medycal** equipment consist of | 42 | 197 |
| | QWERTYCharSub | what is durable medical equipment **xonsist** of | 42 | 182 |
| Naturality | RemoveStopWords | ~~**what is**~~ durable medical equipment consist ~~**of**~~ | 37 | 199 |
| | T5DescToTitle | ~~**what is**~~ durable medical equipment ~~**consist of**~~ | 35 | 136 |
| Ordering | RandomOrderSwap | **medical** is durable **what** equipment consist of | 43 | 200 |
| Paraphrasing | BackTranslation | what is **sustainable** medical equipment ~~**consist of**~~ | 23 | 93 |
| | T5QQP | what is durable medical equipment ~~**consist of**~~ | 26 | 105 |
| | WordEmbedSynSwap | what is durable **medicinal** equipment consist of | 27 | 124 |
| | WordNetSynSwap | what is **long lasting** medical equipment consist of | 16 | 71 |

# Experiments

## Method

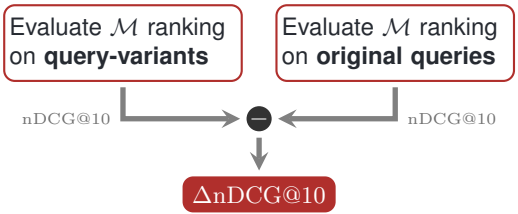
### Ranking robustness

### Embedding robustness

# Experiments

## Method

### Ranking robustness

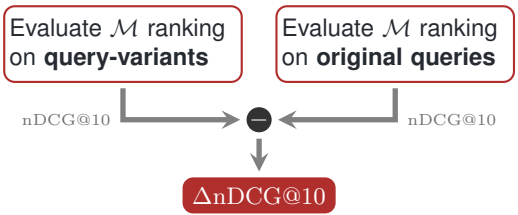



### Embedding robustness

# Experiments

## Method

### Ranking robustness




### Embedding robustness

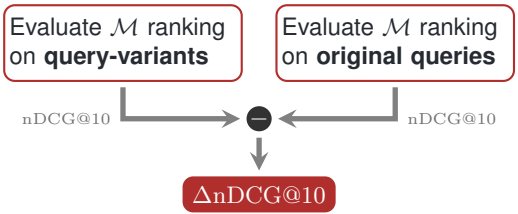
### Note

- Ideally, $\Delta\text{nDCG@10}$ is 0
- $\Delta\text{nDCG@10} > 0$ means $\mathcal{M}$ is more effective on the query variant

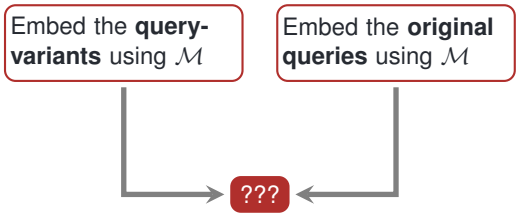
# Experiments

## Method

### Ranking robustness




### Embedding robustness

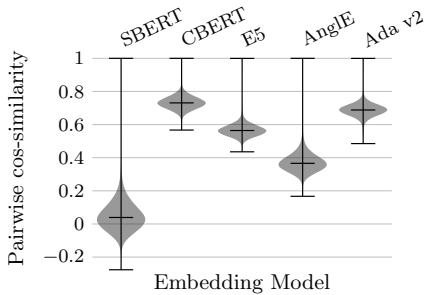



### Note

- Ideally, $\Delta\text{nDCG@10}$ is 0
- $\Delta\text{nDCG@10} > 0$ means $\mathcal{M}$ is more effective on the query variant

# Experiments

## Anisotropy in Embedding Models

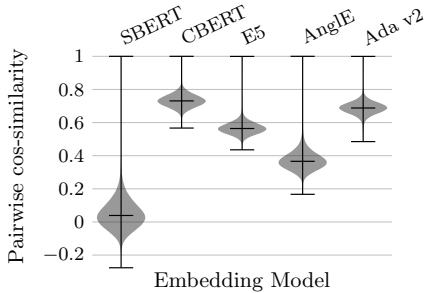


# Experiments

## Anisotropy in Embedding Models

- High cossim $\not\Rightarrow$ semantically similar

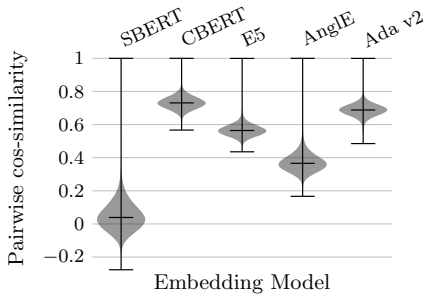
  *(Unrelated inputs have a cossim of $0.71$ for CBERT)*



# Experiments

## Anisotropy in Embedding Models

- High cossim $\not\Rightarrow$ semantically similar
  *(Unrelated inputs have a cossim of $0.71$ for CBERT)*
- Embeddings are not uniformly distributed
  *(“Anisotropic”)*

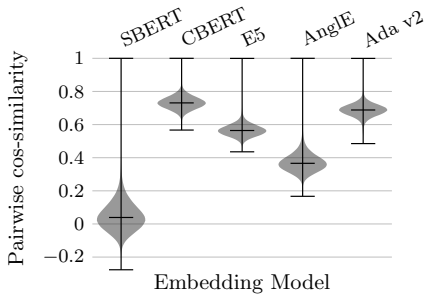


# Experiments

## Anisotropy in Embedding Models

- High cossim $\not\Rightarrow$ semantically similar
  *(Unrelated inputs have a cossim of $0.71$ for CBERT)*
- Embeddings are not uniformly distributed
  *(“Anisotropic”)*

▶ Cossim can’t be compared across models



# Experiments

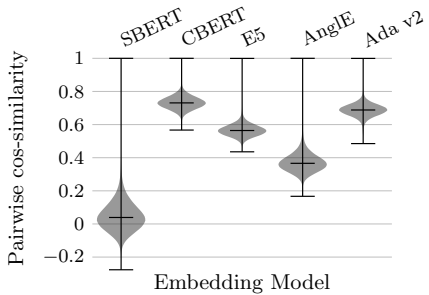
## Anisotropy in Embedding Models

- High cossim $\not\Rightarrow$ semantically similar
  *(Unrelated inputs have a cossim of $0.71$ for CBERT)*
- Embeddings are not uniformly distributed
  *("Anisotropic")*

- Cossim can't be compared across models
- Adjust cossim for anisotropy

$$\text{adjcossim}(v, v') = \frac{\text{cossim}(v, v') - \mu}{1 - \mu}$$

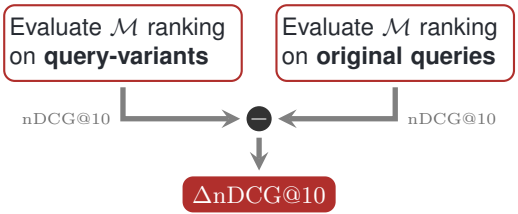
*Expected cossim for two arbitrary inputs*



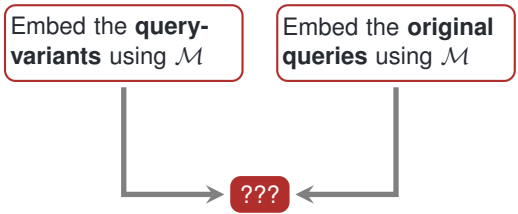
# Experiments

## Method

### Ranking robustness




### Embedding robustness



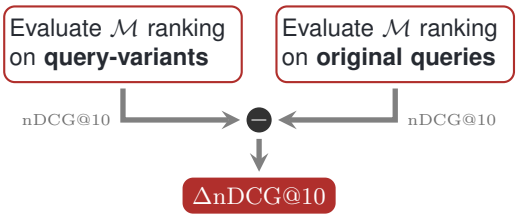

### Note

- Ideally, $\Delta\text{nDCG@10}$ is 0
- $\Delta\text{nDCG@10} > 0$ means $\mathcal{M}$ is more effective on the query variant

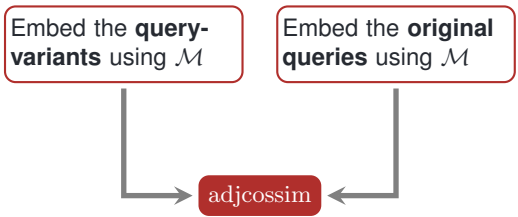
# Experiments

## Method

### Ranking robustness




### Embedding robustness



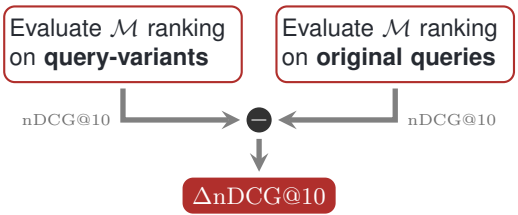

### Note

- Ideally, $\Delta\text{nDCG@10}$ is 0
- $\Delta\text{nDCG@10} > 0$ means $\mathcal{M}$ is more effective on the query variant

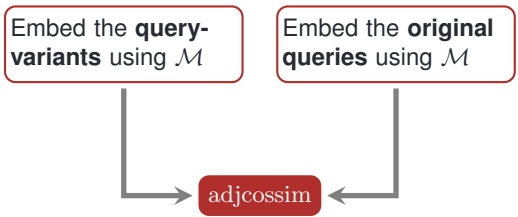
# Experiments

## Method

### Ranking robustness




### Embedding robustness




### Note

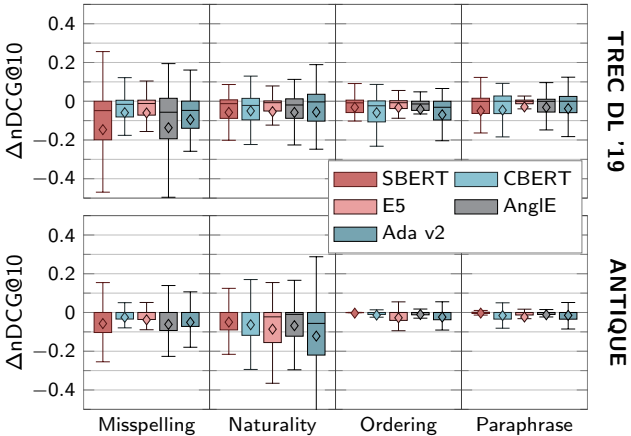
- Ideally, $\Delta\text{nDCG@10}$ is 0
- $\Delta\text{nDCG@10} > 0$ means $\mathcal{M}$ is more effective on the query variant

### Note

- Ideally, $\text{adjcossim}$ is 1
- The expected $\text{adjcossim}$ of two arbitrary inputs is 0

# Results

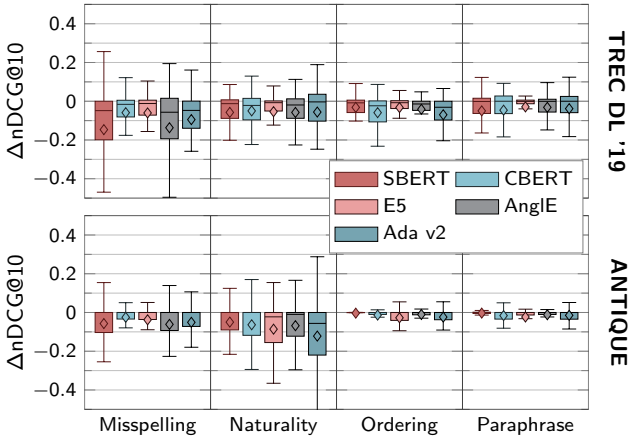
## Ranking Robustness



# Results

## Ranking Robustness

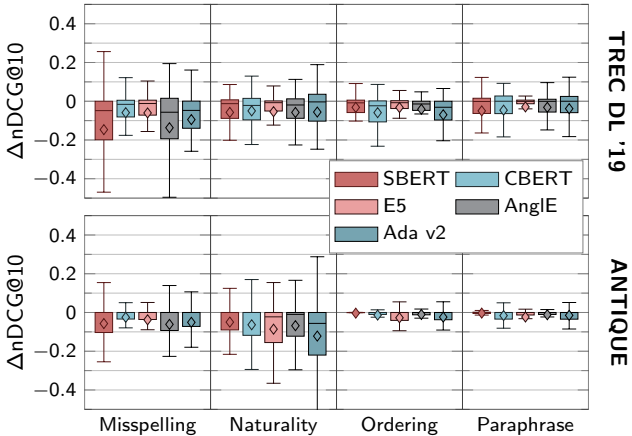
- $\Delta\text{nDCG@10}$ sometimes positive but mostly negative



# Results

## Ranking Robustness

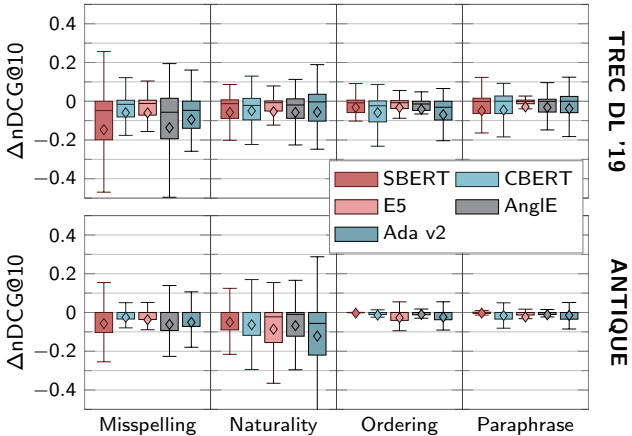
- $\Delta\mathrm{nDCG@10}$ sometimes positive but mostly negative
- Only effectiveness degradation is statistically significant



# Results

## Ranking Robustness

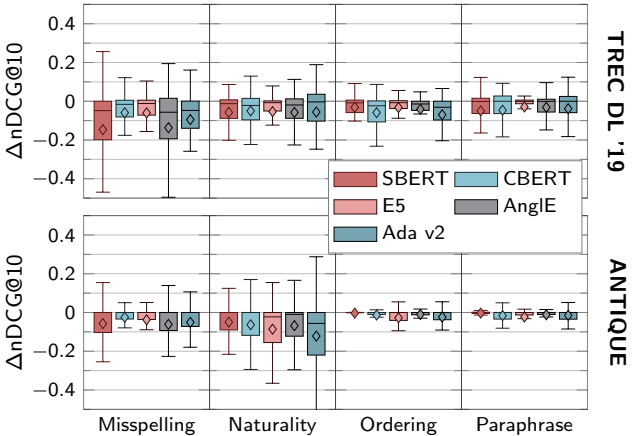
- $\Delta\text{nDCG@10}$ sometimes positive but mostly negative
- Only effectiveness degradation is statistically significant
- Smaller spread on ANTIQUE (except for naturality)



# Results

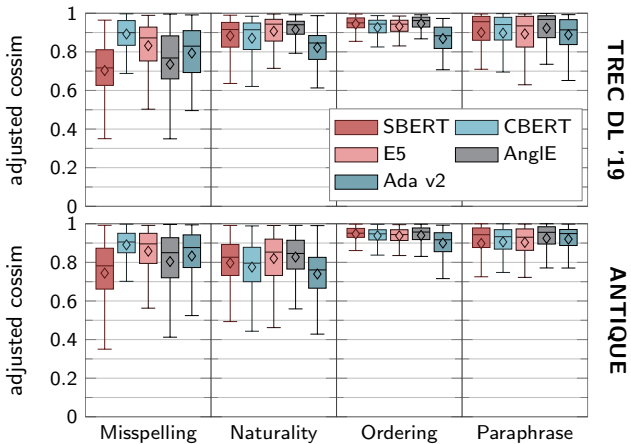
## Ranking Robustness

- $\Delta$nDCG@10 sometimes positive but mostly negative
- Only effectiveness degradation is statistically significant
- Smaller spread on ANTIQUE (except for naturality)
- On ANTIQUE, all models are least robust to naturality



# Results

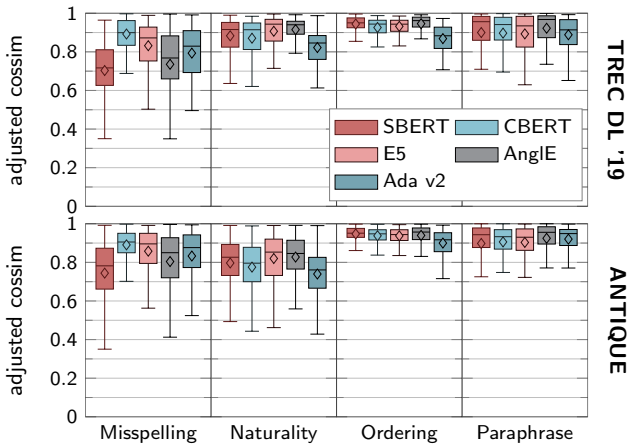
## Embedding Robustness



# Results

## Embedding Robustness

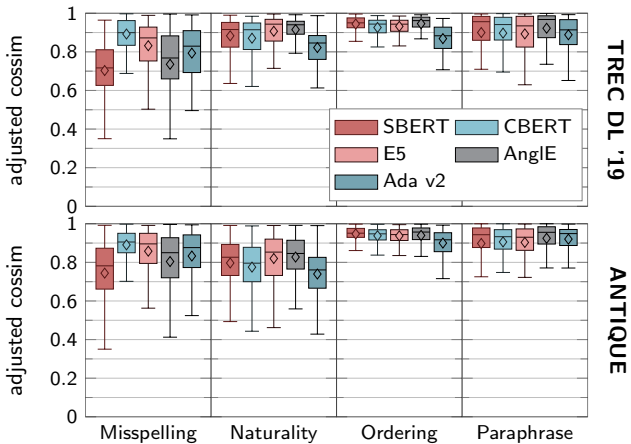
- Ordering and paraphrasing the easiest



# Results

## Embedding Robustness

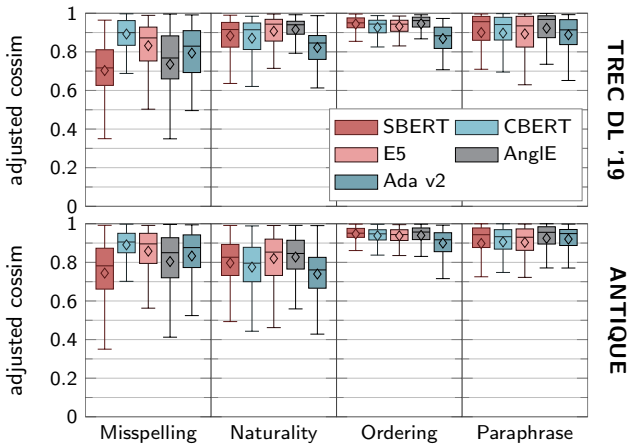
- Ordering and paraphrasing the easiest
- CBERT the most robust to typos



# Results

## Embedding Robustness

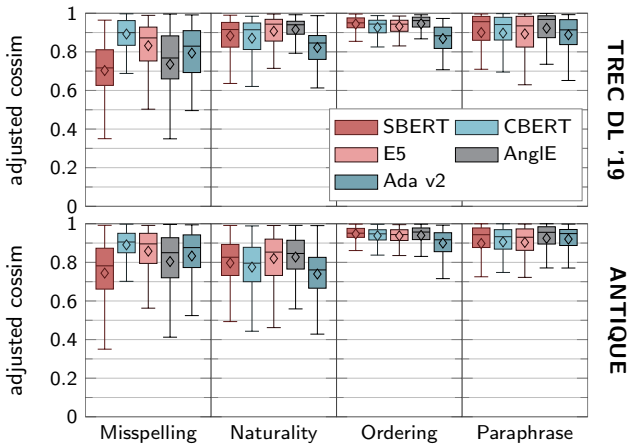
- Ordering and paraphrasing the easiest
- CBERT the most robust to typos
- AnglE the most robust except to typos



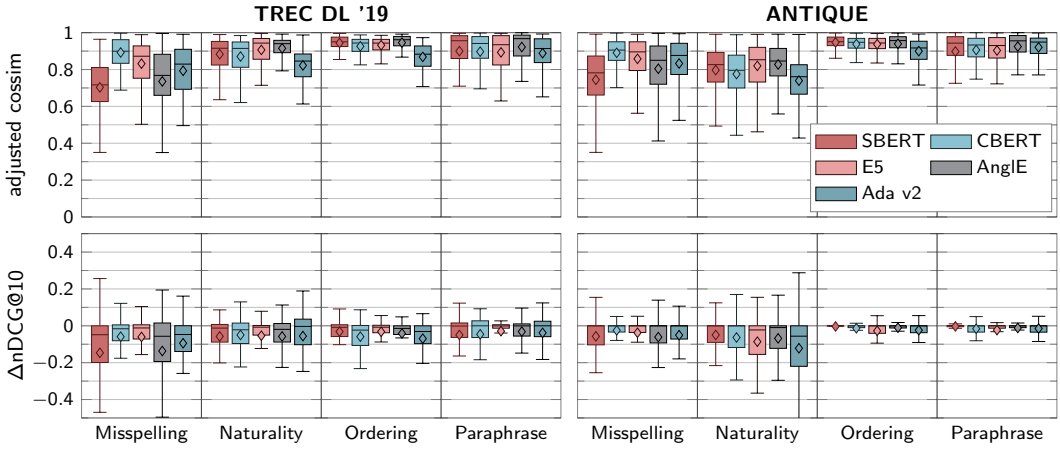
# Results

## Embedding Robustness

- Ordering and paraphrasing the easiest
- CBERT the most robust to typos
- AnglE the most robust except to typos
- E5 Mistral in median similarly robust to the most robust model (but larger spread)



# Results




None of the models are robust

# Additional Experiments 1

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

# Additional Experiments 1

## Prompting Robustness

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

**Research Question**

Can robustness simply be prompted?

# Additional Experiments 1

## Prompting Robustness

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

Author instruction

**Instruction** (excerpt)

*Given a web search query, retrieve relevant passages that answer the query*

**Research Question**

Can robustness simply be prompted?

# Additional Experiments 1

## Prompting Robustness

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

Author instruction

**Instruction** (excerpt)

*Given a web search query, retrieve relevant passages that answer the query*

*Given a web search query, fix typos and retrieve relevant passages that answer the query*

**Research Question**

Can robustness simply be prompted?

# Additional Experiments 1

## Prompting Robustness

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

Author instruction

**Instruction** (excerpt)

- *Given a web search query, retrieve relevant passages that answer the query*
- *Given a web search query, fix typos and retrieve relevant passages that answer the query*
- *Synthesize the ideal query to express the given informationneed and retrieve relevant passages for it*

**Research Question**

Can robustness simply be prompted?

# Additional Experiments 1

## Prompting Robustness

**Note**

E5-Mistral is based on Mistral-7b-instruct and can be prompted via

```
Instruct: instruction
Query: query
```

Author instruction

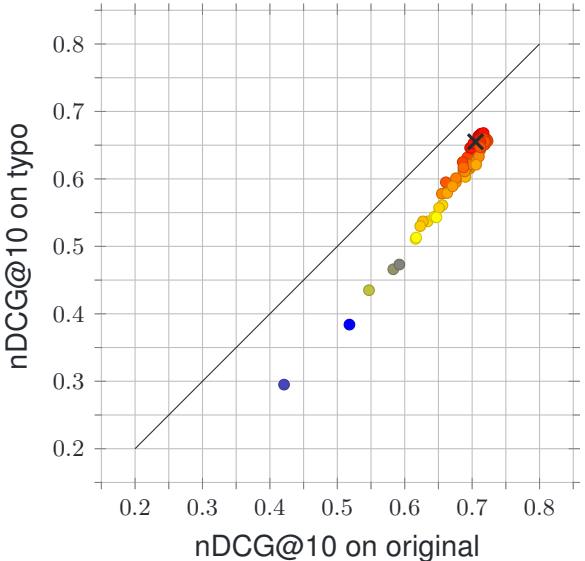
| Instruction (excerpt) |
| --- |
| *Given a web search query, retrieve relevant passages that answer the query* |
| *Given a web search query, fix typos and retrieve relevant passages that answer the query* |
| *Synthesize the ideal query to express the given informationneed and retrieve relevant passages for it* |
| *Do what you want* |

**Research Question**

Can robustness simply be prompted?

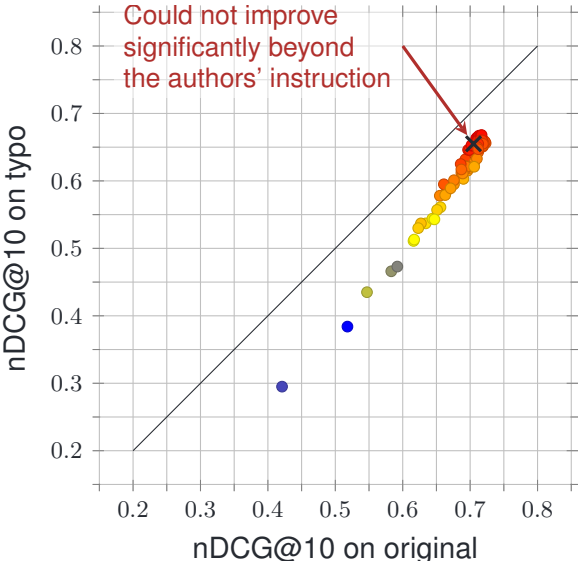
# Additional Experiments 1

## Prompting Robustness



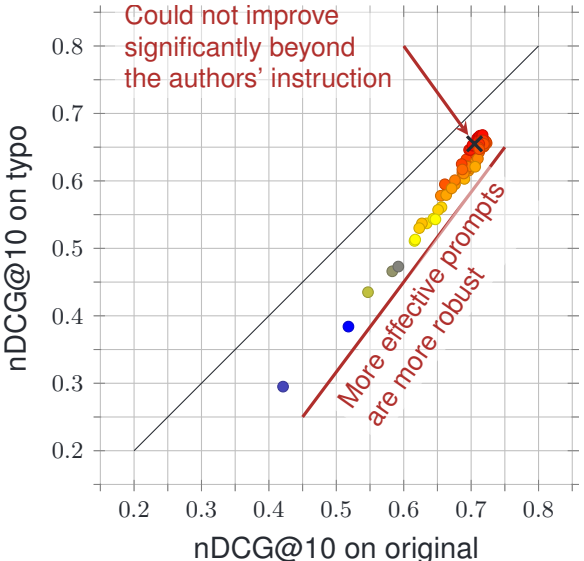
# Additional Experiments 1

## Prompting Robustness



# Additional Experiments 1

## Prompting Robustness



# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

### Setup

- Create a training set using Penha et al.'s transformations

# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

**Setup**

- Create a training set using Penha et al.'s transformations
- Prompt-tune E5-Mistral & Fine-tune CBERT

# Additional Experiments 2

## Training Robustness

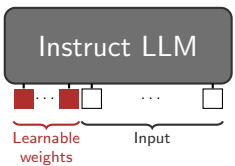
**Research Question**

How does training on more query variations affect robustness?

### Setup

- Create a training set using Penha et al.'s transformations
- Prompt-tune E5-Mistral & Fine-tune CBERT

### Prompt-tuning

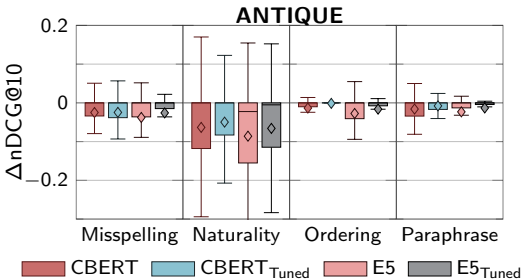


# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?



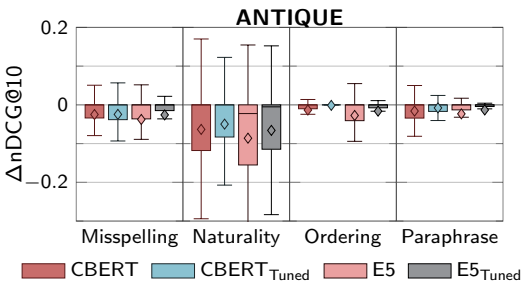
# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

- Both models improved robustness across all categories



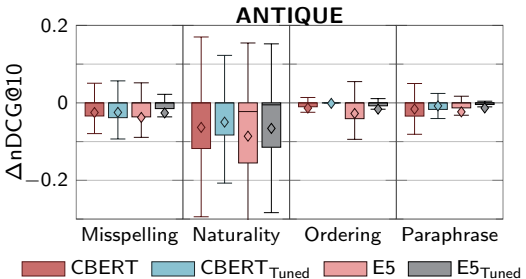
# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

- Both models improved robustness across all categories
- ... but mean effectiveness is not improved and



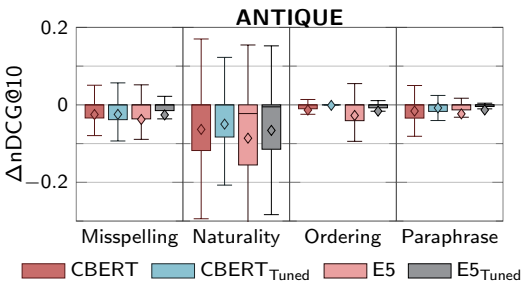
# Additional Experiments 2

## Training Robustness

**Research Question**

How does training on more query variations affect robustness?

- Both models improved robustness across all categories
- ... but mean effectiveness is not improved and
- degradation is statistically significant



# Summary

- Transformer-based embedding models are effective rankers

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations
- We tested embedding models that are...
  - more recent
  - typo-aware
  - larger
  - commercial

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations
- We tested embedding models that are...
  - more recent
  - typo-aware
  - larger
  - commercial
- Result: they, too, are not robust

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations
- We tested embedding models that are...
  - more recent
  - typo-aware
  - larger
  - commercial
- Result: they, too, are not robust
- Prompt- and fine-tuning on a query variation dataset can improve robustness

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations
- We tested embedding models that are...
  - more recent
  - typo-aware
  - larger
  - commercial
- Result: they, too, are not robust
- Prompt- and fine-tuning on a query variation dataset can improve robustness

**Take-away**

- Transformer-based embedding models are still not robust and

tim.hagen@uni-kassel.de
temir.org

# Summary

- Transformer-based embedding models are effective rankers
- ... but not robust to query variations
- We tested embedding models that are...
  - more recent
  - typo-aware
  - larger
  - commercial
- Result: they, too, are not robust
- Prompt- and fine-tuning on a query variation dataset can improve robustness

**Take-away**

- Transformer-based embedding models are still not robust and
- query variation datasets are needed so that typos and keyword queries are not out-of-distribution

tim.hagen@uni-kassel.de
temir.org