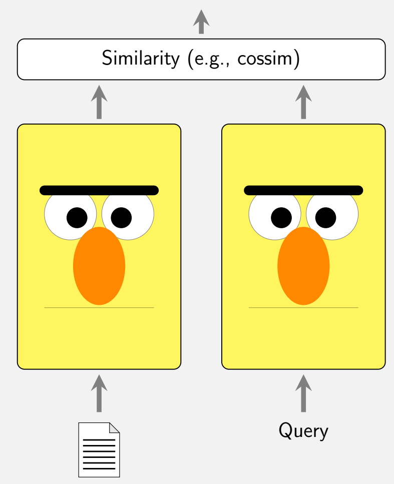
Are transformers robust to query variations?

Revisiting Query Variation Robustness of Transformer Models

Tim Hagen, Harry Scells, and Martin Potthast *(University of Kassel and hessian.AI)*

Problem Statement



1. Compute document embeddings
2. Sort documents by embedding-similarity to the query

- $\approx 70\%$ of information seeking queries use **keywords** and 26% contain **typos**.
- Transformers are used for re-ranking in IR
- ↯ Transformers have been **shown not to be robust** to these variations

RQ: How robust are more recent LLM-based embedding models?

Paper Code

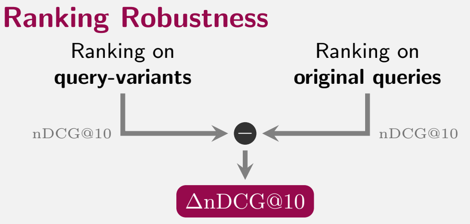
Dataset

Penha et al.'s query variation dataset with **semantically equivalent** query variants

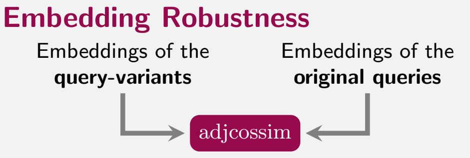
Query variation		Example
Category	Transform. heuristic	
Original		what is durable medical equipment consist of
Misspelling	NeighbCharSwap	what is durable **mdeical** equipment consist of
	RandomCharSub	what is durable **medycal** equipment consist of
	QWERTYCharSub	what is durable medical equipment **xonsist** of
Naturality	RemoveStopWords	~~what is~~ durable medical equipment consist ~~of~~
	T5DescToTitle	~~what is~~ durable medical equipment ~~consist of~~
Ordering	RandomOrderSwap	**medical** is durable **what** equipment consist of
Paraphrasing	BackTranslation	what is **sustainable** medical equipment ~~consist of~~
	T5QQP	what is durable medical equipment ~~consist of~~
	WordEmbedSynSwap	what is durable **medicinal** equipment consist of
	WordNetSynSwap	what is **long lasting** medical equipment consist of

Method

Ranking Robustness



Embedding Robustness



Models

SBERT *(66M parameters)*
Popular embedding model based on DistilBERT$_{\text{Base}}$

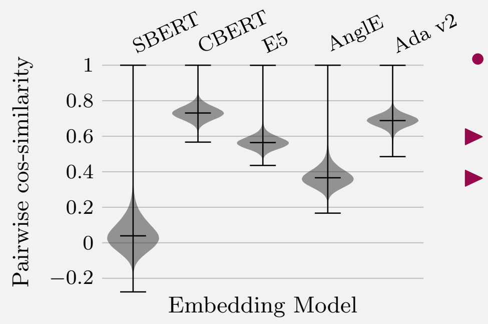
CBERT *(104M parameters)*
Typo-aware variant of BERT$_{\text{Base}}$

E5 *(7B parameters)*
#1 on MTEB; based on Mistral-7B-instruct

AnglE *(335M parameters)*
#2 on MTEB; based on BERT$_{\text{Large}}$

Ada v2
SOTA commercial embedding model

Anisotropy in Embedding Models



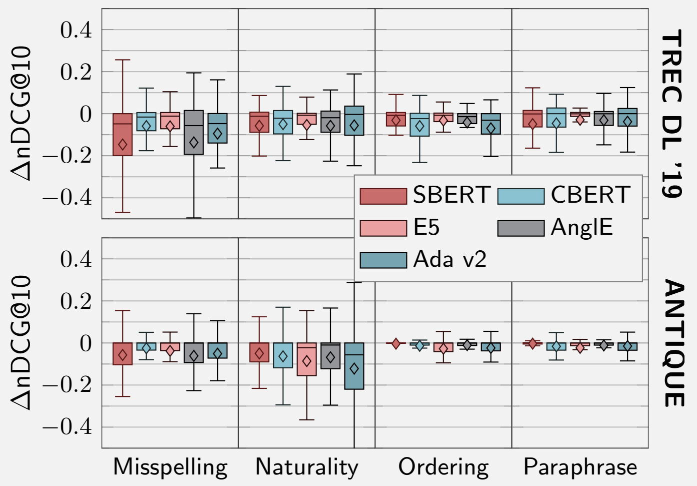
- High cossim $\nRightarrow$ semantically similar *(Mean cossim is 0.71 for CBERT)*
- Embeddings are not uniformly distributed *("Anisotropic")*

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

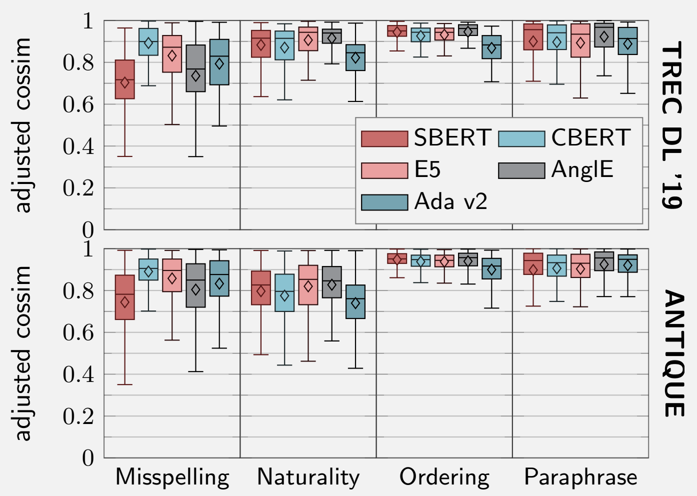
$$\text{adjcossim}(v, v') = \frac{\text{cossim}(v, v') - \mu}{1 - \mu}$$

Expected cossim for two arbitrary inputs

Results



- ΔnDCG@10 is 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



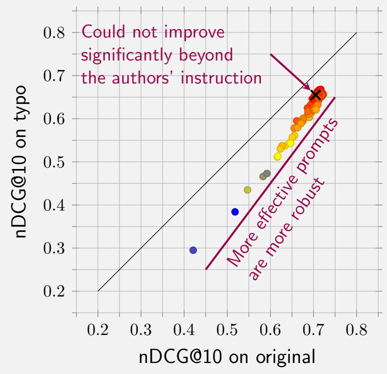
- Ordering and paraphrasing the easiest
- CBERT the most robust to typos
- AnglE the most robust except to typos
- E5 Mistral in median similarly robust to the most robust model (but larger spread)

Additional Experiment 1

Note E5-Mistral is based on an instruct-LLM and is promptable via

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

RQ: Can robustness simply be prompted?

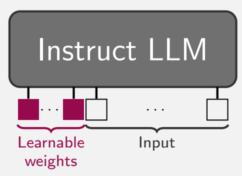


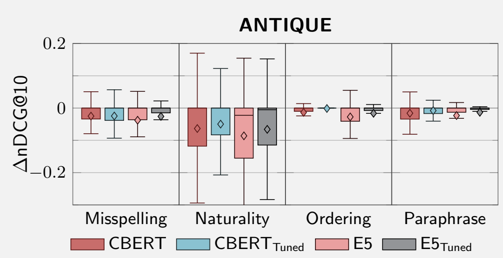
Additional Experiment 2

Prompt tune E5 & fine-tune CBERT for Penha et al.'s transformations.

RQ: How does training on more query variations affect robustness?

Prompt tuning





- Improved robustness across all categories
- ...but still not robust
- statistically significant effectiveness drop