[1] [2] People-Centred AI UNIVERSITY OF SURREY
[3] CENTRE FOR TRANSLATION STUDIES UNIVERSITY OF SURREY

Centrality-aware Product Retrieval and Ranking

Hadeel Saadany[3]
Swapnil Bhosale[2]
Samarth Agrawal[1]
Diptesh Kanojia[2]
Constantin Orăsan[3]
Zhe Wu[1]

EMNLP 2024

Index

01 Introduction

Introduction

E-commerce users often enter **ambiguous and/or complex queries** which often lead to a **mismatch** between the *user's intent* and *retrieved product titles* or documents.

Challenge

- Existing retrieval models primarily rely on **lexical and semantic similarity**.
    a. Ambiguous queries like 'iphone 13', or 'i5 pc 1tb 16gb 8gb gpu' *can lead to many variants*.
    b. Embeddings based approaches do not tackle *repetition of words* from query, present in negative titles.
        i. {**query**: "iphone 13", **title**: "iphone 13 cover white"}
    c. Existing approaches fails to capture the **user's true intent**, *particularly for alphanumeric patterns*.

Contribution

- Introduction of **User-intent Centrality Optimization (UCO)**
    a. Fine-tunes ranking models to prioritize buyer-centric titles using a dual loss-based approach.
- Curation of **novel evaluation sets** for challenging query-title pair subsets.
- **Significant and consistent improvements** observed in ranking performance across several metrics.

02 Motivation

Ambiguous Queries

Ambiguous product queries (e.g., queries like *"iphone 13", "thomas sabo charm"*) often produce spurious results with a few relevant items on top.

- iPhone 13 is the top result but **2nd result is iPhone 12 Pro**

- "Thomas Sabo charm" query produces results for both charm and bracelet.

Thomas sabo charms with 18k Rose gold pearl

Thomas Sabo charm club bracelet with detachable dragonfly charm

Alphanumeric Queries

Alphanumeric product queries (e.g., model numbers like *"i5 pc 1tb 16gb 8gb gpu"*) often *contain characters and numbers* that hold specific meaning, like denoting a product variant (color, type, model, etc.). However, most search algorithms fail to capture nuances, leading to:

- **Ambiguity:** Variants of the same product get mixed up, like different smartphone models or laptop configs.
- **User Dissatisfaction:** Irrelevant or imprecise results reduce user engagement and satisfaction.

S2716DG

Dell S2716DG LED with G Sync 27" QHD Wide 1440p Gaming Monitor

UP2716D

Dell UP2716D 27" IPS LCD UltraSharp 2K QHD 2560 x 1440 Monitor

Challenges

Impact on E-commerce Platforms

- Lower conversion rates due to irrelevant results.
- Poor user retention as customers may turn to competitors.
- Increased need for manual filtering, adding friction to the user experience.

03

Dataset and Annotation

Graded relevance scores

Query: seahawks metcalf on field jersey L

5 - Perfect: D.K. **Metcalf** #14 Seattle Seahawks Men's **onField** Jersey Navy Blue Size **L**

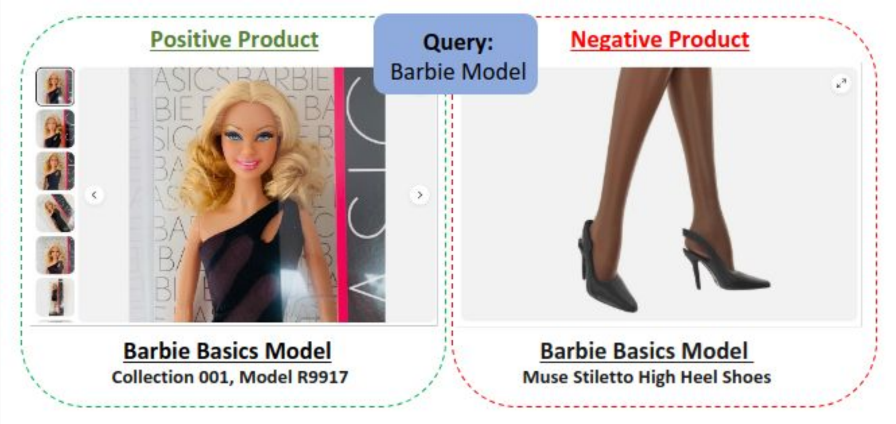
4 - Excellent: D.K. Metcalf Seahawks Men's onField Jersey Navy Blue Size **XL**

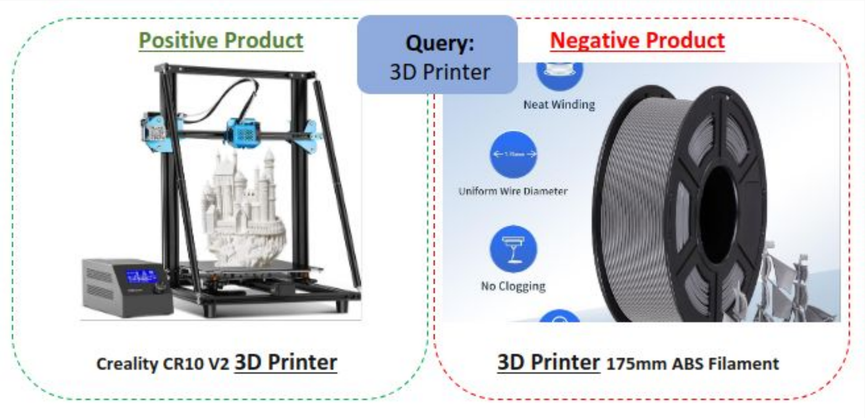
3 - Good: Metcalf Seahawks Men's Jersey Navy Blue Size L

2 - Fair: NWT Devon Witherspoon Seattle Seahawks Blue Jersey Men's Size XL

1 - Bad: Seattle Seahawks Hat 47 Clean Up Adult One Size Adjustable Blue

Centrality


Positive Product
Query:
Barbie Model
Negative Product
Barbie Basics Model
Collection 001, Model R9917
Barbie Basics Model
Muse Stiletto High Heel Shoes


Positive Product
Query:
3D Printer
Negative Product
Neat Winding
Uniform Wire Diameter
No Clogging
Creality CR10 V2 3D Printer
3D Printer 175mm ABS Filament

Dataset and Annotation Process

Curated from eBay's human judged internal datasets, containing real user search queries and corresponding product titles.

Annotation Labels

- **Relevance Ranking (PEGFB schema)**
  - Products are ranked based on how relevant they are to the search query.
- **Centrality Score (Binary)**
  - Indicates whether a product title aligns directly with the user's intent (central or non-central).

Eval Split	# Corpus	# Dev-Q	# Test-Q
CQ	187469	5776	17325
CQ-balanced	46561	5776	17325
CQ-common-str	12508	2117	6351
CQ-alphanum	162115	4111	12333

Challenging Evaluation Sets

- From an existing interal graded relevance dataset, we curate
  - **Common Queries (CQ)** - with relevance >=3 and <3 for harder samples.
    - We balance the data by removing samples from this set -> **CQ-balanced**
  - **CQ-common-str** - Evaluation split with instances which contains query terms within product title too.
  - **CQ-alphanumeric** - Evaluation split with instances which contain alphanumeric patterns in query.

04

User-intent Centrality Optimization

Methodology for Improving Product Ranking

Models

1. **eBERT**
    - a variant of BERT which has been pre-trained on eBay item titles alongside the Wikipedia corpus
2. **eBERT-Siam**
    - A Siamese network variant designed to generate similar embeddings for product titles, enhancing relevance calculation
    - Trained jointly with query and tiles with cosine similarity as matching function

User-Intent Centrality Optimization (UCO)

- Fine-tunes models to prioritize product titles central to user intent, especially for complex alphanumeric queries.

User-Intent Centrality Optimization (UCO)

We fine-tune both models to recognize the most central product titles, ensuring that products closely matching the user's search intent are ranked higher.

Optimization Task

- The model performs **binary classification** for each query-title pair instance in the data. The task is based on our hypothesis that *modelling for centrality based on user-intent reflected within the query should improve product ranking*.

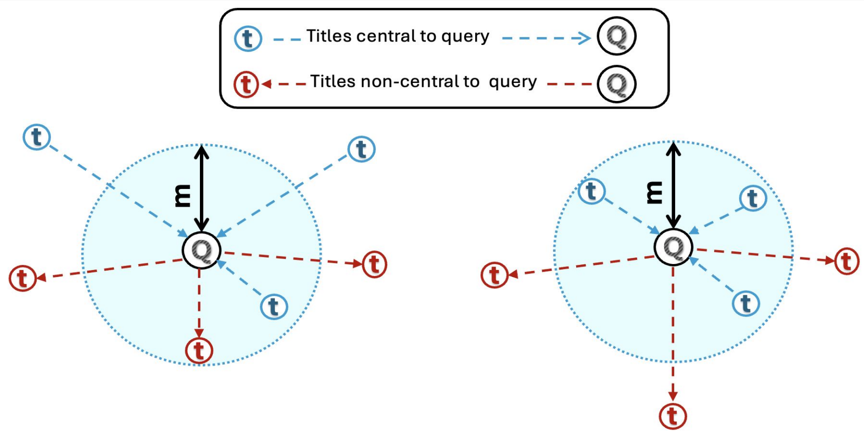
- We combine **Multiple Negative Ranking Loss (MNRL)** and **Online Contrastive Loss (OCL)** to optimize the ranking for harder query-product pairs. This ensures the model focuses on challenging cases where *intent is harder to detect*.

Multiple Negative Ranking Loss (MNRL)

$$\text{MNRL} = \sum_{i=1}^{P} \sum_{j=1}^{N} max(0, f(q, p_i) - f(q, n_j) + margin)$$

Encourages the model to reduce the distance between a query and relevant product titles (positives) while increasing the distance from irrelevant titles (negatives).

Works with multiple negative samples, giving the model better context to learn nuanced differences in relevance.

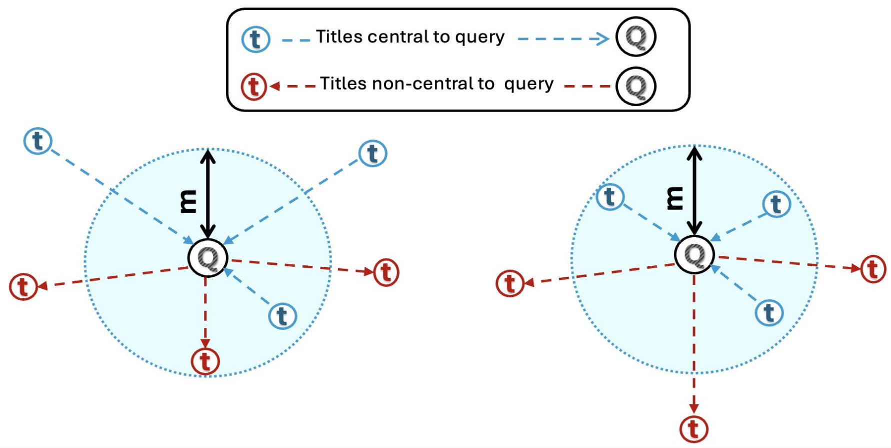


Online Contrastive Loss (OCL)

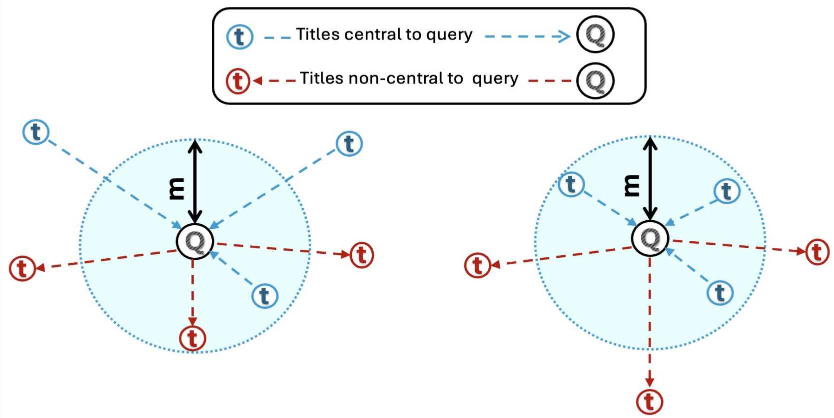
$$\mathrm{OCL} = Y * D + (1\text{-}Y) * \max(\mathrm{margin}\text{-}D,\ 0)^2$$

Focuses on hard cases—pairs of query and product titles where the relevant products are far away (hard positives) and irrelevant ones are too close (hard negatives).

Only optimizes these difficult cases, making the model more precise in handling tricky query-product matches.



Dual-Loss Optimization: Ablation Test



Loss	eBERT		eBERT-siam	
	NDCG@5	MRR@10	NDCG@5	MRR@10
MNRL	0.7139	0.7899	0.7254	0.8016
OCL	0.5497	0.6559	0.5812	0.6978
MNRL + OCL	0.7488	0.8189	0.7698	0.8347

Dual-loss optimization is used to improve the model's ability to distinguish between relevant and irrelevant product titles.

MNRL outperforms OCL in all cases, however, the **combination always fares better.**

05

Results

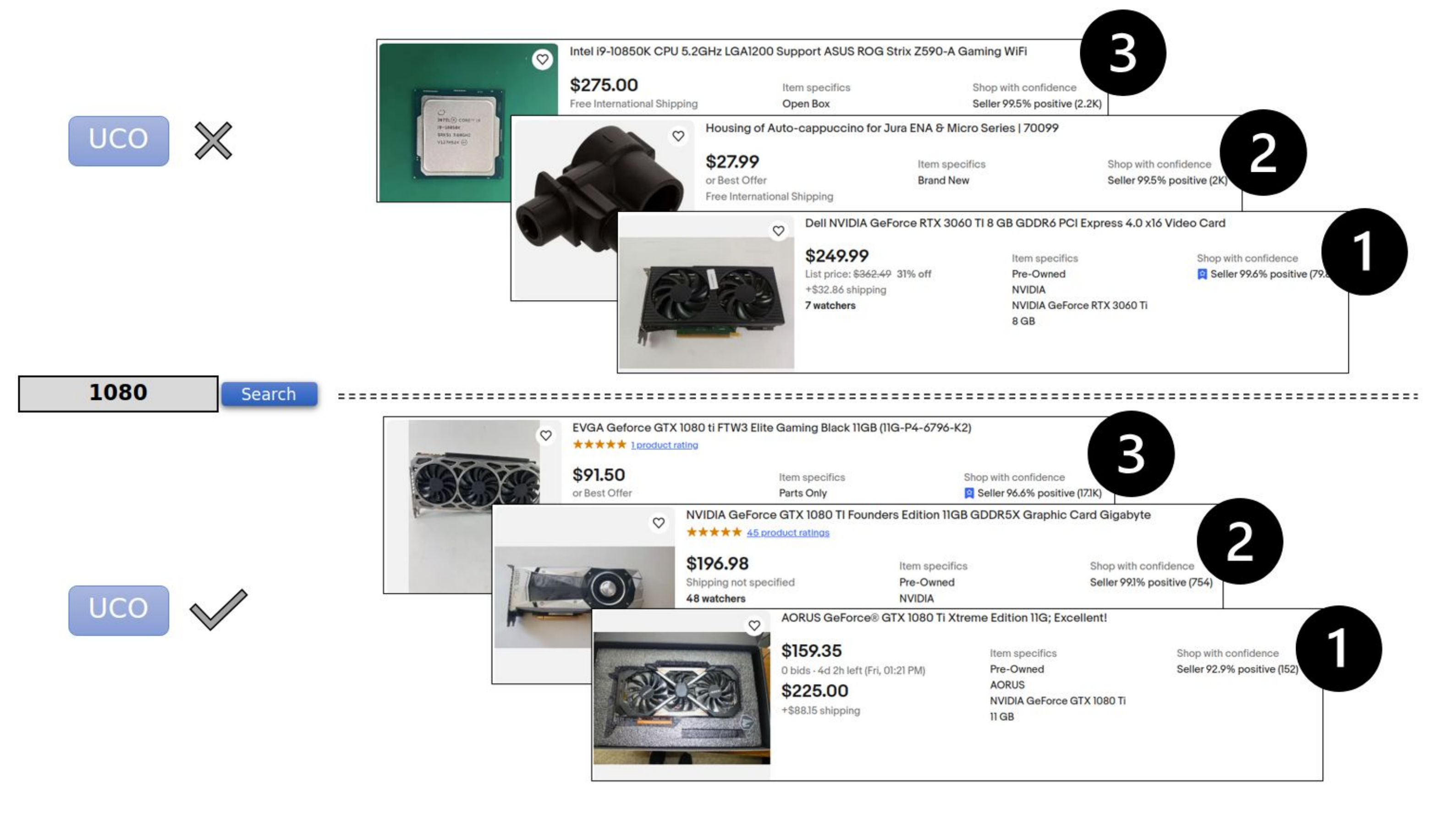
Results

Over the hard baseline (eBERT-siam):

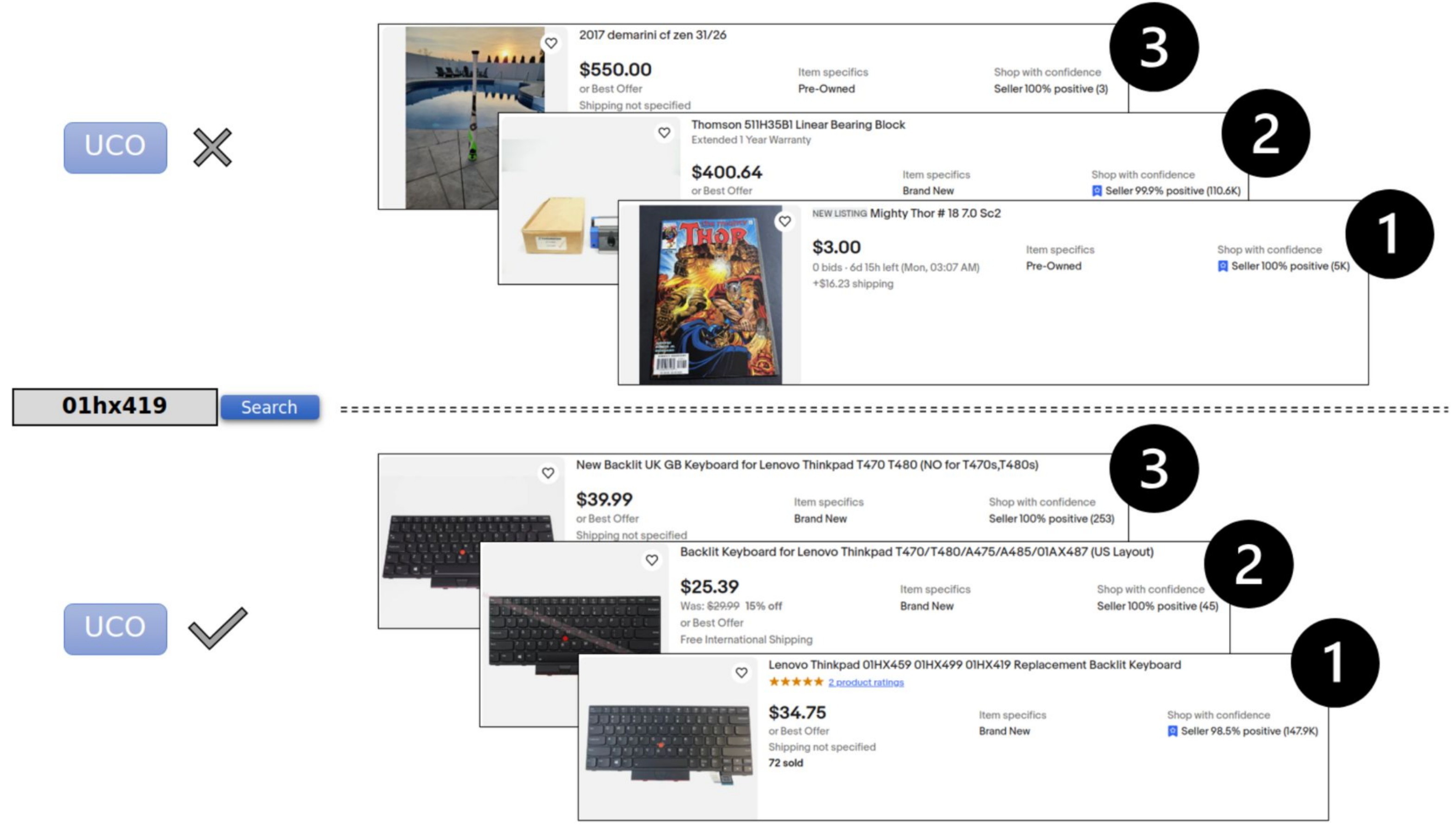
- **On CQ-common-str**, UCO enhances ranking performance across all metrics, with improvements up to **11.3% in NDCG@3** and **7.3% in Recall@3**, highlighting UCO's strength in distinguishing relevant items despite repeated terms.
- **On CQ-alphanum**, UCO achieves consistent gains across metrics, with **10.9% improvement in NDCG@3** and **11.5% in Recall@3**, showing its effectiveness in handling complex product identifiers with high precision.

Encoder	UCO	Precision@k (↑) 3	5	10	Recall@k (↑) 3	5	10	NDCG@k (↑) 3	5	10	MRR (↑) @10
***CQ* test**											
BERT	✗	16.20	13.03	8.93	11.31	14.41	18.83	0.1912	0.1818	0.1833	0.2771
eBERT	✗	20.71	17.25	12.54	14.46	19.19	26.26	0.2392	0.2330	0.2430	0.3415
	✓	64.76	55.74	39.22	49.63	63.92	79.65	0.7439	0.7488	0.7672	0.8189
eBERT (siam)	✗	55.25	48.33	34.90	42.36	56.09	72.22	0.6315	0.6428	0.6704	0.7263
	✓	66.25	57.16	40.20	51.18	65.79	81.66	0.7635	0.7698	0.7886	0.8347
***CQ-balanced* test**											
BERT	✗	7.13	4.94	2.95	21.26	24.58	29.33	0.1824	0.1961	0.2115	0.1862
eBERT	✗	9.72	6.94	4.22	29.02	34.58	42.07	0.2428	0.2657	0.2899	0.2495
	✓	28.57	18.15	9.50	85.40	90.42	94.62	0.7851	0.8059	0.8197	0.7789
eBERT (siam)	✗	25.99	16.68	8.89	77.66	83.08	88.59	0.6888	0.7112	0.7291	0.6784
	✓	29.19	18.39	9.58	87.26	91.58	95.43	0.8046	0.8225	0.8351	0.7965
***CQ-common-str* test**											
BERT	✗	9.41	6.31	3.65	28.15	31.47	36.35	0.2532	0.2669	0.2828	0.2579
eBERT	✗	12.62	8.64	5.00	37.79	43.10	49.92	0.3272	0.3491	0.3714	0.3315
	✓	32.03	19.58	9.92	95.84	97.65	98.87	0.9091	0.9166	0.9206	0.8979
eBERT (siam)	✗	29.93	18.76	9.68	89.57	93.58	96.50	0.8194	0.8361	0.8456	0.8063
	✓	32.12	19.64	9.92	96.11	97.94	98.93	0.9117	0.9193	0.9226	0.9003
***CQ-alphanum* test**											
BERT	✗	20.54	16.65	11.47	13.45	17.32	22.82	0.2333	0.2176	0.2226	0.3350
eBERT	✗	23.35	19.54	13.77	15.53	20.76	27.85	0.2630	0.2516	0.2617	0.3739
	✓	64.58	57.27	40.35	44.05	59.97	77.00	0.7119	0.7094	0.7344	0.8018
eBERT (siam)	✗	60.67	54.10	38.54	41.32	57.10	74.20	0.6652	0.6654	0.6951	0.7618
	✓	67.10	59.70	41.81	46.07	62.72	79.76	0.7375	0.7371	0.7609	0.8171

Qualitative Analysis



Qualitative Analysis



06

Conclusion and Future Work

Conclusion and Future Directions

- Our approach shows **a substantial improvement in product retrieval performance** across all metrics.
    a. Evident regardless of the backbone encoder employed
    b. Consistent across metrics

- Proposed **dual-loss based optimisation,** helps the model identify hard negatives, *i.e.*, semantically relevant but non-central titles.

- Our approach is **product domain-agnostic** for query types which contain a substring within the title pair, the performance on which we will evaluate in near future.

- In future, we aim to expand query and product aspects with explanations assisted by an LLM to generate explanations for computing relevance.

Thank you

Questions?

References

1. Fredrik Carlsson,Amaru Cuba Gyllensten, Evangelia Gogoulou,Erik Ylipää Hel- Iqvist, and Magnus Sahlgren. 2020. Semantic re-tuning with contrastive tension. In International conference on learning representations.

2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

3. Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652 (2017).

4. Po-Sen Huang,Xiaodong He,Jianfeng Gao,Li Deng,Alex Acero,and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2333–2338.