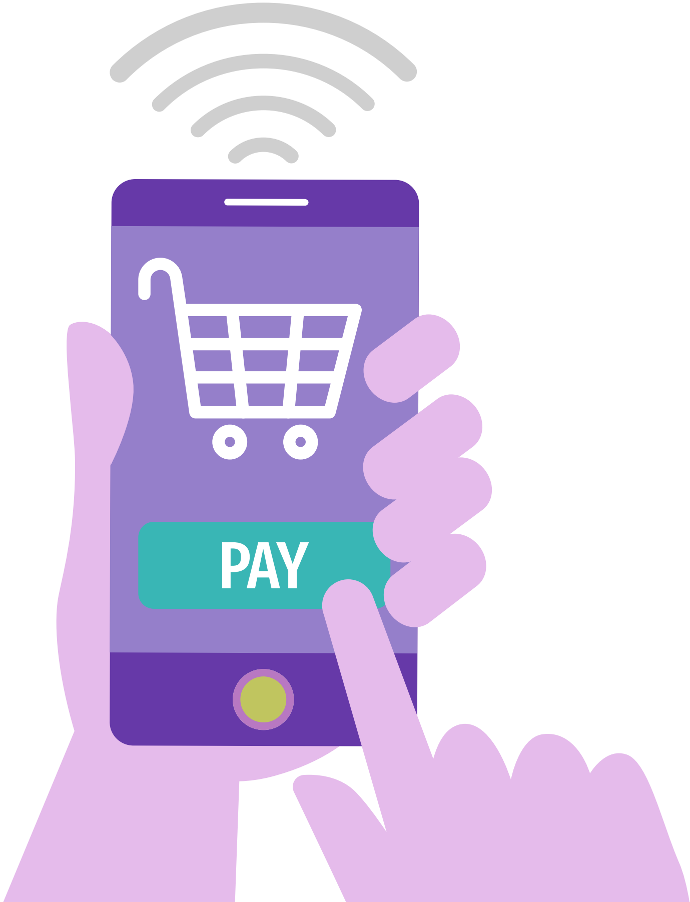
Improving Search Relevance using Amazon ESCI Dataset

Sohani Patki sp8110
Sruthi Raghavan sr7892

References:1.,2.

Problem Statement



- E-commerce search engines often return irrelevant or weakly relevant results
- Poor search relevance → lower user satisfaction & sales
- Goal: Improve query–product relevance prediction using ML models

Dataset - Amazon ESCI

Amazon ESCI (Exact, Substitute, Complement, Irrelevant)	Query–product pairs with relevance labels	Multi-class relevance classification problem	Real-world e-commerce search data
Exact	Substitute	Complement	Irrelevant

References: 2.,3.

Project Objective

Analyze ESCI dataset characteristics

Train a **teacher model** to predict relevance

Improve search ranking quality

Establish a baseline for **future knowledge distillation**

References:1

Removed missing records

Retained records belonging to English locale only

Concatenated query and product details and mapped (E,S,C,I) to soft labels (1,0.5,0,0)

Teacher Model

A

Base Model: meta-llama/Llama-3.2-1B model

B

Fine-tuned on Amazon ESCI dataset using LoRA (PEFT)

C

Input: Query+Product Title+Product Description+Product Bullet points+Product Brand+ Product Color

D

Output: Logit score indicating probability that item is relevant to query

References:1.,2.,4.,5.,6

Training Setup

01

Loss Function: BCEWithLogitsLoss

02

Optimizer: AdamW

03

Evaluation Metrics: MSE

04

Training: Models were trained on an NVIDIA A100 GPU with an ~85% training and 15% testing data split.

Teacher Model Fine-tuning Details Summary

- Base model: meta-llama/Llama-3.2-1B
- LoRA Adaptation: train q and v layers (trainable params: 13,631,488)
  - LoRA Rank 128
  - LoRA Dropout 0.05
- Learning rate: 1e-5
- Add score head MLP layer to compute logit score for relevance
- Uses BCEWithLogitsLoss and AdamW optimiser
- Train-test split % (85-15)
- Model trained over 7 epochs, checkpointed and saved upon every 20% epoch completion

References:1.,4.,5.,6.,7.

Results & Observations

Model successfully learns relevance patterns

Provides strong baseline for:

- Student models
- Ranking optimization

MSE for the model on test data (31003 samples) is 0.155861

Conclusion & Future Work

- Trained a reliable teacher model on ESCI data

- Future Work:

   i. Knowledge distillation to smaller models
   ii. Ranking-based loss functions
   iii. Multilingual relevance modeling

Acknowledgements and Citations

1. Shang, Hongwei, et al. "Knowledge Distillation for Enhancing Walmart E-commerce Search Relevance Using Large Language Models." Companion Proceedings of the ACM on Web Conference 2025. 2025.

2. Reddy, Chandan K., et al. "Shopping queries dataset: A large-scale ESCI benchmark for improving product search." arXiv preprint arXiv:2206.06588 (2022).

3.Kaggle Dataset Link: https://www.kaggle.com/datasets/notsalmankhan/amazon-esci-shopping-queries

4.Grattafiori, Aaron, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).

5.Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

Acknowledgements and Citations

6. Mangrulkar, S., et al. "State-of-the-art parameter-efficient fine-tuning methods." 2022,

7. Ilya Loshchilov, Frank Hutter, "Decoupled Weight Decay Regularization", International Conference on Learning Representations (ICLR) 2019. https://arxiv.org/abs/1711.05101