# QUAKER: Query Understanding And top-K Embedding Retrieval

## CIS 5200 Final Project

Xiangjun Fu, Xiaoyang Wan, Dong Dong

University of Pennsylvania

December 2025

# Outline

Introduction

Dataset Analysis

Methodology: QUAKER

Evaluation

Related Work

Conclusion

# Motivation

- **Context**: E-commerce platforms increasingly use natural language interfaces.
- **The Challenge**:
  - Real-world queries are **obscure**, **conversational**, and **under-specified**.
  - Example: ”I need something to fix my wobbly table…” vs. ”Furniture pads”.
  - Keyword-based systems (TF-IDF/BM25) fail to capture these nuances.

- **The Dilemma**:
  - **End-to-end Neural Models**: Accurate but computationally prohibitive on millions of items.
  - **Sparse Retrieval**: Fast but lacks semantic understanding.
- **Our Solution**: **QUAKER**, a dual-stage framework balancing semantic expressiveness with efficiency.

# Problem Definition

**Goal**: Learn a scoring function $s_\theta(q, i)$ to measure semantic relevance between query $q$ and item $i$.

- **Catalog**: $\mathcal{I} = \{i_1, \ldots, i_M\}$ ($M \approx 21,000$ in test set).
- **Data**: Each item $i$ has metadata $m_i$ and category $c_i$.
- **Objective**: Place ground-truth item $i^*$ in the top-$K$ results ($K = 200$).

$$\text{Acc}_K = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}\left[ i^* \in \text{TopK}_{s_\theta}(q) \right]$$

# Dataset Construction

**Source 1: Amazon-C4 Dataset** (McAuley Lab)

- 21,223 user reviews rephrased by ChatGPT into vague queries.
- Contains: `query`, `item_id`.

**Source 2: Sampled Item Metadata**

- 1 Million items sampled from Amazon Reviews 2023.
- Contains: `item_id`, `category`, `metadata`.

**Merging Process**:

- Matched Amazon-C4 queries with Metadata via `item_id`.
- Filtered out items with empty or extremely short metadata ($< 10$ tokens).

# Exploratory Data Analysis

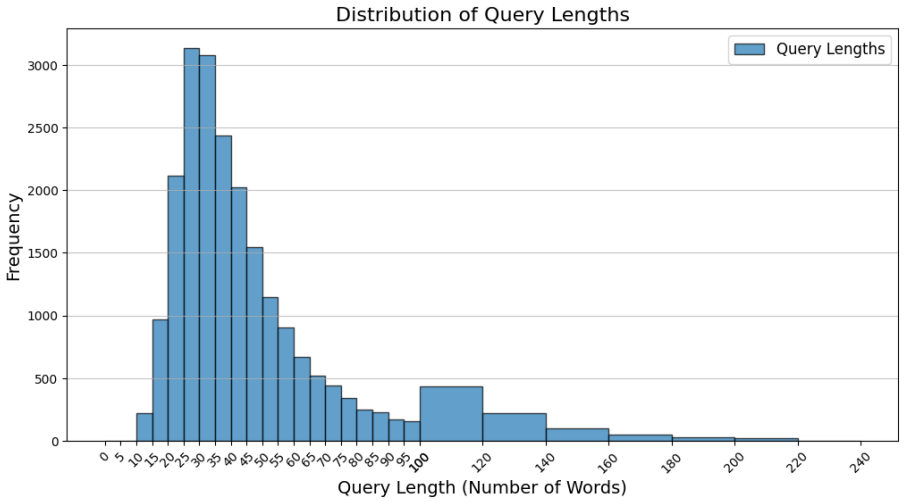



Figure 1: Query Length Distribution

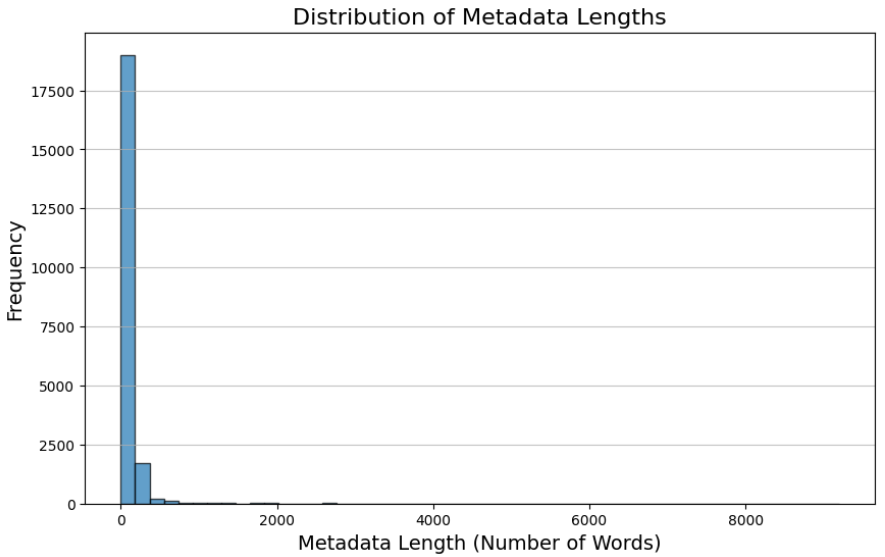
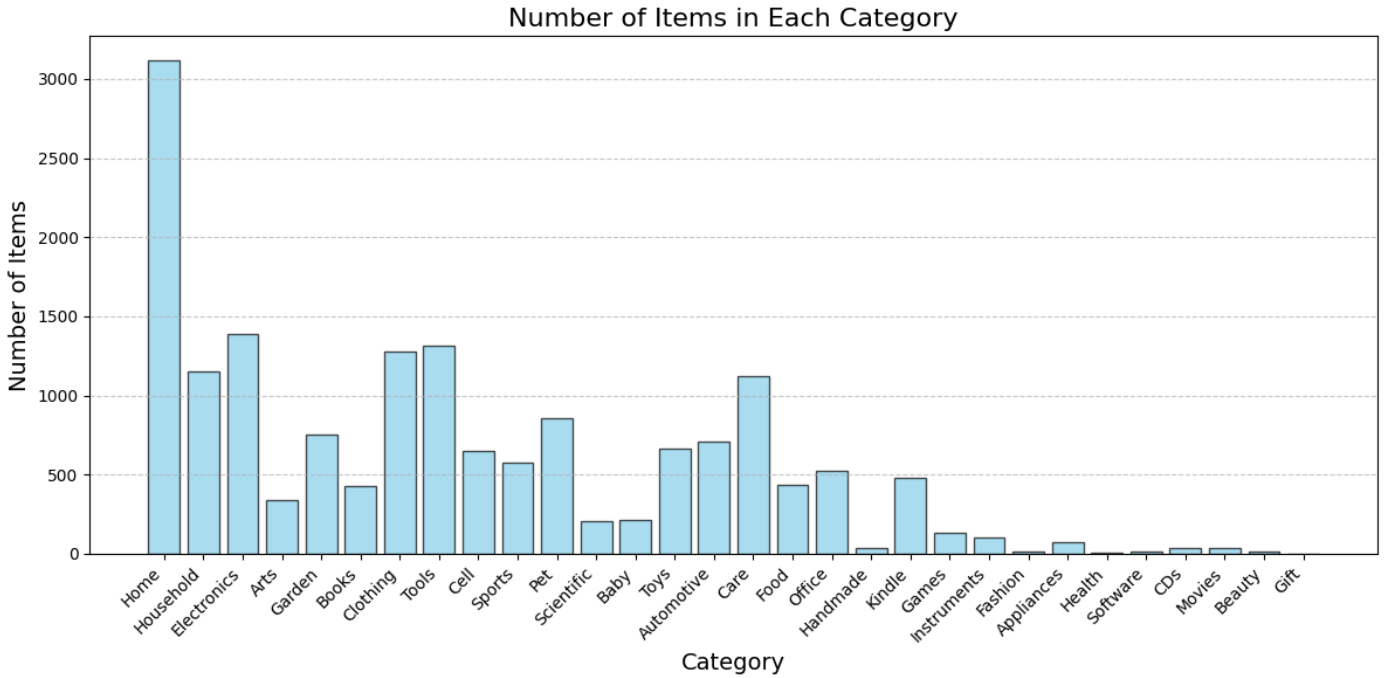



Figure 2: Metadata Length Distribution

- **Queries**: Mostly short (< 220 words), conversational tone.
- **Metadata**: Long-tail distribution; ”short metadata” is a key error source.

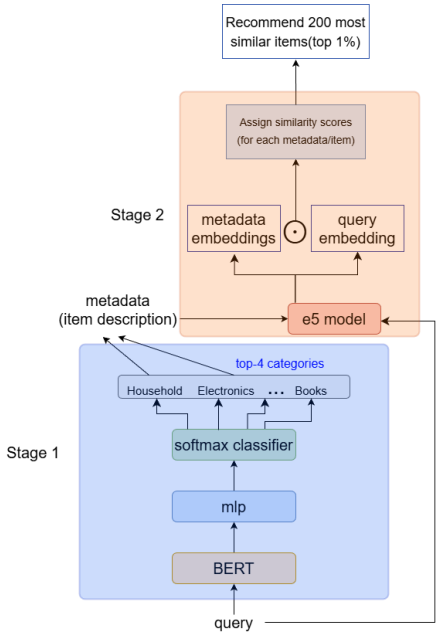
# Category Imbalance




- **Dominant Classes**: ”Home” (>3,000 items), ”Electronics”.
- **Impact**: High priors for dominant classes; challenge for tail categories like ”Gift”.

# QUAKER Framework Overview

**QUAKER**: **Q**uery **U**nderstanding **A**nd top-**K** **E**mbedding **R**etrieval




- **Decomposition**: Splits retrieval into *Coarse Filtering* (Stage 1) and *Fine Ranking* (Stage 2).

# Stage 1: Category Prediction (Query Understanding)

- **Model**: Fine-tuned BERT Classifier (3 epochs).
- **Input**: Raw user query $q$.
- **Output**: Top-4 predicted categories $C_{\text{top}}(q)$.
- **Mechanism**:
  - Restricts search space to $\mathcal{I}(q) = \{i : c_i \in C_{\text{top}}(q)\}$.
  - Eliminates hard negatives from irrelevant categories.
- **Result**: 90.06% accuracy (True category in Top-4).

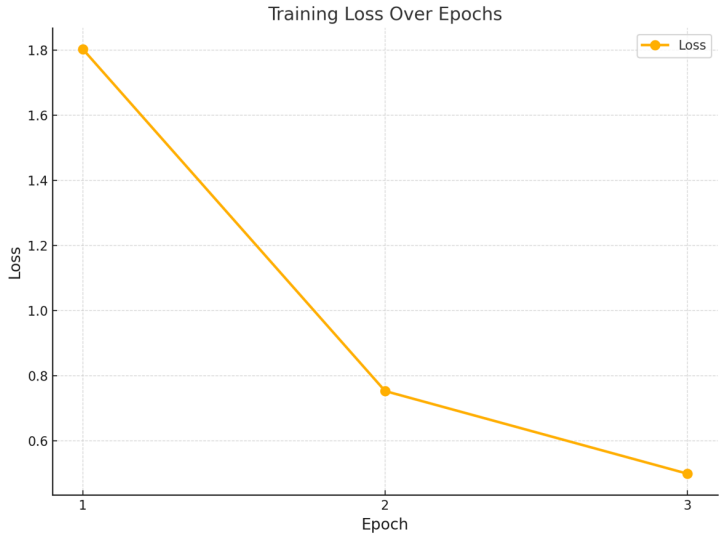
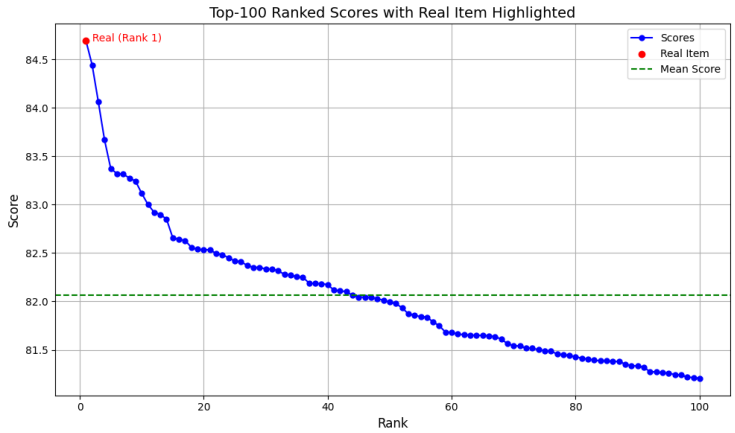



Figure 3: Training Loss

# Stage 2: Semantic Ranking (Dense Retrieval)

- **Model**: E5 (Text Embeddings).
- **Why E5?**: Weakly-supervised contrastive pre-training aligns well with retrieval tasks.
- **Process**:
  1. Encode Query: $v_q = \text{E5}(q)$
  2. Encode Metadata: $v_i = \text{E5}(m_i)$ for all $i \in \mathcal{I}(q)$.
  3. Rank by Cosine Similarity: $\frac{v_q \cdot v_i}{\|v_q\| \|v_i\|}$.
- **Advantage**: Captures semantic matches beyond keyword overlap.



# Quantitative Results (Top-200 Accuracy)

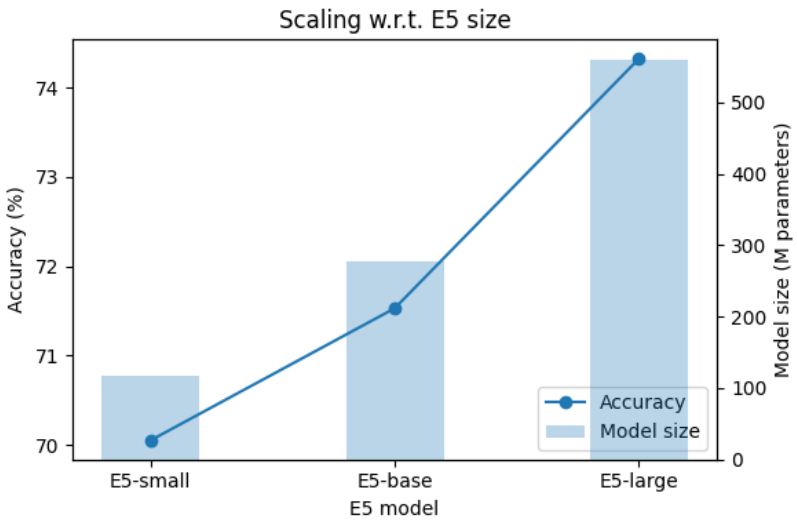
| Configuration | Top-200 Accuracy (%) | | |
|---|---|---|---|
| | E5-small | E5-base | E5-large |
| BERT-base | 69.58 | 71.16 | 73.88 |
| BERT-large | 70.52 | 71.90 | **74.77** |
| TF-IDF Baseline | | 42.43 | |
| No Classifier (Full Retrieval) | | 65.21 | |

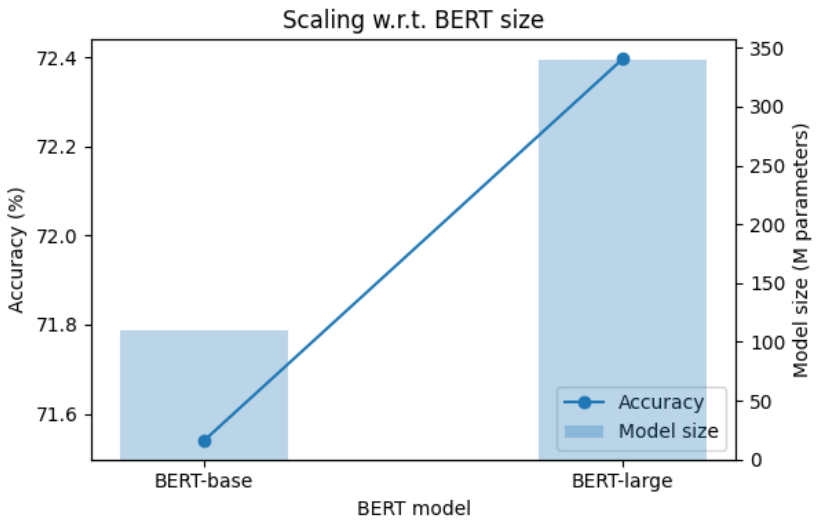
Table 1: Test set performance comparison.

**Key Findings**:

- QUAKER (74.77%) outperforms TF-IDF (42.43%) by **>30 points**.
- "No Classifier" (65.21%) proves Stage 1 is crucial for both accuracy and efficiency.

# Scaling Analysis




- **Monotonic Gains**: Increasing model size (BERT or E5) consistently improves accuracy.
- **Interpretation**: Larger models learn better semantic compression and fine-grained distinctions.

# Error Attribution Analysis

We decompose the system performance:

$$\text{Acc}_{\text{overall}} \approx \text{Acc}_{\text{cls}} \times \text{Acc}_{\text{ret}|\text{cls}}$$

For our best model (BERT-large + E5-large):

- $\text{Acc}_{\text{cls}}$ (Category Prediction): **90.56%**
- $\text{Acc}_{\text{ret}|\text{cls}}$ (Retrieval given correct category): **82.56%**
- $\text{Acc}_{\text{overall}}$: **74.77%**

**Error Sources**:

- **37.4%** of errors are due to **Classification** (Wrong category).
- **62.6%** of errors are due to **Retrieval** (Right category, but wrong ranking).
- **Insight**: Improving the Retriever offers the largest potential gain.

# Qualitative Case Study I

**Query**: ”I am looking for shoes that are super comfy... paired with professional attire... perfect height... neutral color...”

**Top Results**:

1. **Rank 1**: ”Ryka Women’s Devotion Plus 2 Walking Shoe... exceptional cushion... neutral color...” (*Highly Relevant*)
2. **Rank 2**: ”Valentine Gnomes Plush...” (*Irrelevant - likely hard negative*)

**Observation**:

- The model often retrieves semantically relevant items even if they are not the specific ground-truth item labeled in the dataset.
- Suggests potential for **Dataset Enrichment** using model predictions.

# Related Work

- **Dense Retrieval**:
  - DPR (Karpukhin et al.), ColBERT (Khattab et al.).
  - Shift from lexical (BM25) to semantic matching.
- **Multi-Stage Architectures**:
  - ”Retrieve-and-Rerank” paradigm to balance speed and accuracy.
  - QUAKER adapts this by using Category Prediction as the ”Retrieve” (filter) stage.
- **E-commerce Query Understanding**:
  - Specialized models like BLaIR and TransRec bridge the gap between user reviews and item metadata.

# Conclusion

- **Summary**: QUAKER effectively handles obscure e-commerce queries using a scalable two-stage approach.
- **Impact**:
  - **Efficiency**: Reduces candidate pool via high-accuracy category prediction.
  - **Performance**: 74.77% Top-200 accuracy vs 42.43% baseline.
- **Key Takeaway**: Explicit query understanding (categorization) significantly amplifies the power of dense retrieval models.
- **Future Work**:
  - Refine dataset with model-assisted labeling.
  - Explore larger LLMs for query rewriting/enrichment.

# Thank You!

Q & A