Database-Augmented Query Representation for Information Retrieval

Anonymous ACL submission

Abstract

Information retrieval models that aim to search for documents relevant to a query have shown multiple successes, which have been applied to diverse tasks. Yet, the query from the user is oftentimes short, which challenges the retrievers to correctly fetch relevant documents. To tackle this, previous studies have proposed expanding the query with a couple of additional (user-related) features related to it. However, they may be suboptimal to effectively augment the query, and there is plenty of other information available to augment it in a relational database. Motivated by this fact, we present a novel retrieval framework called Database-Augmented Query representation (DAQu), which augments the original query with various (query-related) metadata across multiple tables. In addition, as the number of features in the metadata can be very large and there is no order among them, we encode them with the graph-based set-encoding strategy, which considers hierarchies of features in the database without order. We validate our DAQu in diverse retrieval scenarios, demonstrating that it significantly enhances overall retrieval performance over relevant baselines.

1 Introduction

Information Retrieval (IR) is the task of fetching query-relevant documents from a large corpus. Traditional approaches have focused on sparse retrieval, which searches for documents that yield the highest lexical match with the query (Robertson et al., 1994). Recently, neural language models have led to the introduction of dense retrieval models, which represent both the query and the document in a learnable latent space and then calculate their similarity on it (Karpukhin et al., 2020; Izacard et al., 2022; Chen et al., 2024a). Notably, these IR methods have gained much attention in the era of Large Language Models (LLMs), due to their ability to assist LLMs help generating accurate answers with evolving knowledge from an external source (Cho et al., 2023; Jeong et al., 2024).

Despite such a huge advantage of IR in NLP, it faces a critical challenge that information captured in a query itself is oftentimes not sufficient to retrieve its relevant documents, due to the scarcity of information within its (shorter) text. To tackle this challenge, previous work has focused on enriching representations of queries or documents by expanding them with additional texts or augmenting their representation spaces (Jeong et al., 2022; Jagerman et al., 2023; Lin et al., 2023a). Yet, despite their improvements, those approaches are still limited in that they rely on the capability of models themselves (e.g., LLMs) used during augmentation, though there can be external knowledge sources (for augmentation) associated with the user query (such as the user's purchase history for shopping).

While some other work has considered these additional sources, enhancing the representation of queries with them, they leverage only a single source of information stores, especially the one specific to the user (who issues the query) (Gupta et al., 2019; Zhang et al., 2020; Deng et al., 2021; Buss et al., 2023). However, in the real-world, data (including queries) is usually mapped into the database and linked to other data within it, which means that plenty of information that can be potentially used for query enrichment is available on the relational database (Fey et al., 2023). For example, online platforms like e-commerce often use relational databases to store and link structured information such as user profiles, purchase histories, and prior interactions. Similarly, healthcare databases connect patient queries to records like medical histories and lab results, while travel databases associate queries with itineraries and customer profiles.

Therefore, in this work, we introduce a novel IR paradigm, Database-Augmented Query representation (DAQu), which augments representations of queries by searching for and connecting their associated information across multiple tables within

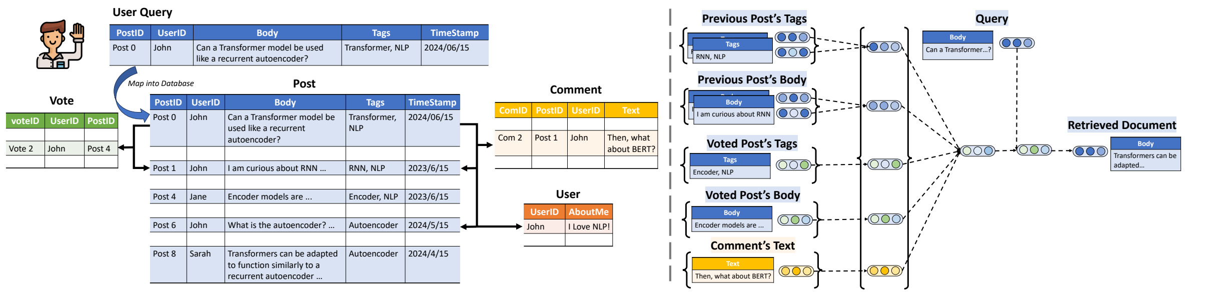

Figure 1: A conceptual illustration of our proposed DAQu, which shows a link among multiple tables for the given query (Left) and visualizes a graph-based set-encoding strategy that encodes metadata hierarchically for query augmentation (Right).

the relational database. As shown in Figure 1, consider the task of identifying the answer post that the user would most likely to vote as the best. In this scenario, we can not only represent the query with its own information but also with its relevant information within and across the multiple tables. Specifically, we can use metadata in the same table, such as its tags, but also metadata spread over other multiple tables, which include user-specific information, such as previous posts, answers (that they voted for), bios, and badges (that they earned). For example, given the question from the user, "Can a Transformer model be used like a recurrent autoencoder?", user tags like "Transformer" and "Autoencoder" can emphasize the focus on these specific concepts. Further, the user's past questions about "RNNs" and "Autoencoders" reveal an existing familiarity with these topics, while the Vote table highlights which answers the user has previously favored, offering further insight into their preferences. However, the volume of these metadata can be extremely large, and simply expanding the query with additional terms in the metadata (as done in existing query expansion work (Gupta et al., 2019; Deng et al., 2021)) is not feasible due to the limited context length of LMs. Moreover, since there is no inherent order for the elements in the metadata, the query augmentation approach should ensure order invariance when incorporating this information.

To this end, we further propose to encode various query-related metadata within and across multiple tables over the relational database, based on a graph set-encoding scheme. Specifically, this strategy models metadata for query expansion as a two-layer hierarchical graph structure, and, within this structure, the first layer aggregates query-related elements (cells) within each column into a column-level representation, and next the second layer aggregates these column-level representations into a query-level representation. For example, consider a query from the Stack Exchange dataset in Figure 1, which is linked to metadata such as the user's profile, previous posts, and associated tags. Then, each individual attribute (e.g., a tag, a user bio, and a body of the previous post) is first encoded independently. After that, within each column (e.g., tags), these encoded attributes are aggregated to create the column-level representation (e.g., all tags combined into a single vector). Lastly, all column-level representations (for tags, user bios, and previous post content) are aggregated into the final query-level metadata representation that is used to enrich the original query representation. It is worth noting that those two-layer structures (aggregation on column- and query-level) can be viewed as a two-layer graph neural network (Kipf and Welling, 2017; Gilmer et al., 2017) since the first layer models interactions within columns (i.e., intra-column relationships) and the second layer models interactions across columns (inter-column relationships).

We then validate our DAQu on seven different retrieval tasks designed with multiple databases of Stack Exchange, Amazon Product Catalog, and H&M (Fey et al., 2023; Robinson et al., 2024). The experimental results show significant improvements of our DAQu in retrieval performance compared to other query augmentation baselines across diverse scenarios. Moreover, we demonstrate that the graph set-encoding technique operationalized in our DAQu effectively represents metadata, enhancing the representations of queries for retrieval.

2 Related Work

Retrieval In response to a query from a user, the retrieval task is to search for the most relevant documents from a large corpus (such as Wikipedia) (Zhu et al., 2021). Typically, it can be performed with two types of models: sparse and dense retrievers. Specifically, sparse retrievers such as TF-IDF or BM25 (Robertson et al., 1994) represent the query and document based on their terms and frequencies in a sparse vector space, whereas dense retrievers use a trainable dense vector space to embed the query and document usually with language models (Karpukhin et al., 2020; Izacard et al., 2022; Chen et al., 2024a). Recently, due to the limitation of sparse retrievers that are vulnerable to the vocabulary mismatch problem, dense retrieval is widely selected as a default choice and many advancements have been made on it (Ding et al., 2024). For example, DPR (Karpukhin et al., 2020) is a supervised dense retriever with a dual-encoder architecture that is trained discriminatively on the labeled pair of a query and its relevant documents to achieve higher similarity scores than the pair of the query-irrelevant documents. Also, Contriever (Izacard et al., 2022) utilizes a self-supervised learning strategy, which generates its training samples by creating positive pairs from query-related contexts within and across documents, rather than relying on explicitly annotated data. Yet, using only the information within a query for retrieval can be suboptimal, due to the scarcity of information on it.

Query Augmentation for Retrieval Some studies have proposed augmenting the original query with additional information to enhance the retrieval performance (Carpineto and Romano, 2012; Azad and Deepak, 2019). Specifically, traditional augmentation methods have focused on utilizing a lexical knowledge base such as the WordNet (Miller, 1992) to expand the original queries (Bhogal et al., 2007; Zhang et al., 2009). In addition, some other work has implemented statistical models such as RM3 (Jaleel et al., 2004a), which add new terms to the query extracted from the top documents in the initial search results and then adjust their weights based on their importance (Lavrenko and Croft, 2001; Jaleel et al., 2004b; Lv and Zhai, 2009). However, they have been shown to be not very effective and, in some cases, even degraded the performance (Nogueira et al., 2019; Jeong et al., 2021). Therefore, recent work has turned to leveraging neural models to extract or generate query-relevant terms and then append such terms to the original query (Esposito et al., 2020; Zheng et al., 2020; Mao et al., 2021), or exploiting multiple fields within the document itself (Li et al., 2025). Moreover, some studies further use recent LLMs to utilize their remarkable capabilities in generating such terms (Wang et al., 2023b; Shao et al., 2023; Buss et al., 2023; Jagerman et al., 2023; Feng et al., 2024; Dhole and Agichtein, 2024; Xia et al., 2024). However, despite the fact that the query is represented and leveraged in the latent space with the recent dense retrievers, existing work focuses on explicitly expanding its text (instead of manipulating this query representation for augmentation). This approach may be problematic if there is a significant amount of data available to augment the query across multiple relational tables over the database.

Retrieval with Database A natural way to store a collection of data is to use a relational database, that is designed to effectively manage, retrieve, and manipulate data for various applications (Johnson et al., 2016; Fey et al., 2023). To utilize the data in the database, the task of retrieving the tabular structures and the information in them has received much attention. Specifically, some studies have developed the approach to retrieve tables themselves (relevant to the given query) from a large table corpus (Herzig et al., 2021; Wang et al., 2022). Some other work extends this approach, extracting or generating the answer for the query from the retrieved tables (Pan et al., 2021, 2022; Lin et al., 2023b). However, since some real-world questions require multiple tables, recent studies have made further progress, proposing to incorporate multiple tables during retrieval (Kweon et al., 2023; Chen et al., 2024b) or reading the tables (Pal et al., 2023). However, unlike all the aforementioned work that has focused on retrieving the tables themselves and finding relevant cells within them, our work is completely different, which aims to effectively handle the query for document retrieval by using the query-related information spread across multiple tables, to augment the representation of the query.

3 Method

3.1 Preliminaries

We begin with preliminaries, providing formal descriptions of the retrieval and query augmentation.

Dense Retrieval Let us define the query as q and its relevant document as $d \in \mathcal{D}$, where $\mathcal{D}$ is a corpus. To operationalize retrieval, we should be able to calculate the similarity between q and d: $f(q, d)$, where f is a scoring function. Following the bi-encoder architecture for dense retrieval, we obtain the similarity by representing the query and document with encoders Enc_q and Enc_d parameterized by θ_q and θ_d, respectively, formalized as follows:

$$f(q, d) = \text{sim}(\boldsymbol{q}, \boldsymbol{d}),$$
$$\boldsymbol{q} = \mathsf{Enc}_q(q; \theta_q) \quad \text{and} \quad \boldsymbol{d} = \mathsf{Enc}_d(d; \theta_d), \tag{1}$$

where $\boldsymbol{q}$ and $\boldsymbol{d}$ are the query and document representations, respectively. In addition, sim is a similarity metric (e.g., cosine similarity). It is worth noting that the objective of the dense retrieval function f is to rank the pair of query q and its relevant document d^+ highest among all the other pairs with irrelevant documents $\{d_i^-\}_{i=1}^N$. To reflect this, we formalize the training objective, as follows:

$$l = -\log \frac{e^{f(q,d^+)}}{e^{f(q,d^+)} + \sum_{i=1}^N e^{f(q,d_i^-)}}. \quad (2)$$

Query Augmentation for Retrieval To improve the effectiveness of the dense retrieval (while tackling the limited contextual information within the query q), the textual query itself or its representation $\boldsymbol{q}$ can be enriched by augmenting it with the information that is not present in the original q. In this work, to effectively incorporate diverse pieces of information into the query without their order variance, we turn to augmenting the query representation $\boldsymbol{q}$ over the latent space, as follows: $\tilde{\boldsymbol{q}} = \lambda \boldsymbol{q} + (1-\lambda)\boldsymbol{q}'$, where $\tilde{\boldsymbol{q}}$ is the reformulated representation, $\boldsymbol{q}'$ is the representation of the additional information helpful to enrich the original query $\boldsymbol{q}$, and $\lambda \in [0,1]$ is for giving weight to it.

3.2 Database-Augmented Query Representation

We now introduce our Database-Augmented Query representation (DAQu) framework for IR.

Relational Database As a vast amount of information is typically stored in a relational database, we aim to augment the representations of queries with the relevant information within this database. The relational database can be defined as a set of tables: $\mathcal{T} = \{T_i\}_{i=1}^N$, and each table is comprised of a collection of rows $T = \{r_j\}_{j=1}^K$, where N is the number of tables and K is the number of rows.

Note that one of the valuable characteristics of the relational database is that some rows in tables are connected with others in other tables, which facilitates relational linkages and ease of data retrieval. Formally, each row r_i in the table consists of a primary key column that uniquely identifies each row within the table, (potentially) some foreign key columns that link to primary keys in other tables, and other non-key attribute columns providing additional information about the row. In other words, the relationships between primary and foreign keys connect rows across different tables, and other attribute columns store descriptive information. Formally, if a foreign key column f in table T_i references a primary key column p in T_j, we can represent their relationship as (f_i, p_j). Also, all such relationships between tables can be denoted as $\mathcal{L} = \{(f_i, p_j)\}_{(i,j)}$ where $\mathcal{L} \subseteq T \times T$.

For example, analogous to the Amazon database, let's assume that the table T_{review} includes the primary key column REVIEWID, the foreign key column PRODUCTID, and the attribute column TEXT. Also, the table $T_{product}$ has the primary key column PRODUCTID and the attribute column DESCRIPTION. Lastly, the foreign key column PRODUCTID in T_{review} points to the primary key column in $T_{product}$. Then, the relationships between those two tables can be represented with a pair of primary and foreign keys: $(\text{PRODUCTID}_{review}, \text{PRODUCTID}_{product})$.

Query Augmentation with Relational Database Recall the equation to augment the representation of the given query ($\tilde{\boldsymbol{q}} = \lambda \boldsymbol{q} + (1-\lambda)\boldsymbol{q}'$). Here, $\boldsymbol{q}'$ is the representation that we obtain from the query-related information within the relational database, and we now turn to explain how to get this $\boldsymbol{q}'$.

Formally, each query that the user requests can be considered as one row r_j in a certain table T_i. For example, in the Stack Exchange dataset, the query that the user posts is stored in the table as one row: $r \in T_{post}$, where this row (query) r consists of the primary key (POSTID), the foreign key (USERID), and the multiple attributes (such as BODY, TAGS, and TIMESTAMP). Then, based on the following relational structure of this database:

$$\begin{aligned}\mathcal{L} = \{&(\text{USERID}_{user}, \text{USERID}_{post}),\\ &(\text{USERID}_{vote}, \text{USERID}_{post}),\\ &(\text{POSTID}_{post}, \text{POSTID}_{comment}), ...\},\end{aligned} \quad (3)$$

the row for the query in the post table can be linked to other rows in different tables, for example, the user table, vote table, and comment table connected with USERID and POSTID columns (Figure 1).

This relational structure of the database allows us to utilize diverse pieces of information when enriching the query representation $\boldsymbol{q}$. Specifically, we can not only use the attributes within the columns of the row for the query (such as BODY and TAGS of the post table T_{post}) but also the attributes of associated rows (to the query) from different tables (such as ABOUTME of the user table T_{user} associated with the column USERID). Formally, all the attributes of rows associated with and used to augment the query (q) can be represented as follows:

$$\begin{aligned}\mathcal{A} = &\{r_{i,j} \mid r_i = q\} \cup\\ &\{r_{i,j} \mid q \in T \text{ and } r_i \in T' \text{ and } (T, T') \in \mathcal{L}\} \cup\\ &\{r_{i,j} \mid r_i \in T \text{ and } q \in T' \text{ and } (T, T') \in \mathcal{L}\},\end{aligned} \quad (4)$$

where $r_{i,j}$ is the value of the jth attribute column of the ith row. Then, based on these attributes (the metadata), we derive their representation $\boldsymbol{q}'$ with the encoder: $\boldsymbol{q}' = \mathsf{Enc}_a(\mathcal{A}; \theta_a)$, described below.

Graph-Structured Set-Encoding We now turn to explain how to operationalize the encoding function $\mathsf{Enc}_a(\cdot)$, which should effectively represent the diverse attributes $\mathcal{A}$ (over the relational database) into $\boldsymbol{q}'$, to enrich the original query representation $\boldsymbol{q}$. To accomplish this objective, one possible strategy is to concatenate all the attribute values, and encode the concatenated value with the encoder or append it to the original query (before encoding), following the existing query expansion work (Zheng et al., 2020; Deng et al., 2021; Dhole and Agichtein, 2024). However, these approaches have a couple of limitations. First, due to the large volume of data in the database, the number of attributes related to the query could be large, and it might be infeasible to encode their concatenated text with the encoder (due to its limited context length). Also, attributes do not have an inherent order (i.e., permutation invariant), making it arbitrary to determine the sequence in which they should be concatenated.

To tackle these challenges, we propose encoding attributes ($\mathcal{A}$) with a graph-structured set-encoding strategy, which differs from and indeed extends the prior set-encoding (Zaheer et al., 2017). Specifically, we first encode every attribute value $r_{i,j}$ in $\mathcal{A}$ into $\boldsymbol{r}_{i,j}$ with an attribute encoder: $\boldsymbol{r}_{i,j} = \mathsf{Enc}_r(r_{i,j}; \theta_r)$, and then aggregate a group of encoded attributes according to each column into the single representation with mean pooling as follows: $\boldsymbol{R}_j = \mathsf{MEAN}(\{\boldsymbol{r}_{i,j}\}_{i=1})$, which then captures the representation of each category (or column) of the metadata. After that, we aggregate all these categorical (column-wise) representations into another representation, which represents the overall metadata for the given query as $\boldsymbol{q}' = \mathsf{MEAN}(\{\boldsymbol{R}_j\}_{j=1})$. Note that this dual-layer structure — aggregating at both the column- and query-levels — resembles a two-layer graph neural network (Kipf and Welling, 2017; Gilmer et al., 2017), where each layer functionally captures the interactions between the attributes in the same column first and the columns over different tables next in a hierarchical manner.

Let's consider the scenario in Figure 1, where the goal is to retrieve the answer post most likely to be selected as the best by the user. Then, the query is encoded into $\boldsymbol{q}$, which is further enriched with the metadata representation $\boldsymbol{q}'$ obtained via the proposed graph-structured set-encoding as follows: the metadata ($\mathcal{A}$) includes attributes such as user comments (COMMENT), tags (TAGS), and the user profile (ABOUTME); each attribute is encoded into a column-level representation, e.g., $\boldsymbol{R}_{\text{COMMENT}} = \mathsf{MEAN}(\{\mathsf{Enc}_r(r_{i,\text{COMMENT}})\}_{i=1})$ (and similarly for others); all column-level representations are aggregated into a single query-level representation: $\boldsymbol{q}' = \mathsf{MEAN}(\{\boldsymbol{R}_{\text{COMMENT}}, \boldsymbol{R}_{\text{TAGS}}, \boldsymbol{R}_{\text{ABOUTME}}\})$, which is used to augment the original query representation.

Efficient Training Strategy with Metadata The number of attributes collected from the relational database is sometimes very large for certain queries, and it may be largely inefficient to consider all of them during training. To address this, we introduce a two-stage sample selection strategy to efficiently train the metadata encoder Enc_r and to efficiently obtain the metadata representation $\boldsymbol{q}'$. Specifically, during training, instead of using all attributes in $\mathcal{A}$ for parameter updates, we randomly sample three attributes for each column and use only them to train the metadata encoder. In addition, while we can use all the remaining attributes (without gradients) to obtain the metadata representation along with the representations of three specific attributes for each column (with gradients), using all the remaining attributes may still be time-consuming and may yield the over-fitting issue; thus, we randomly sample some of them and use only them to obtain the representation $\boldsymbol{q}'$. Meanwhile, in the inference step, we utilize all the metadata attributes available.

4 Experimental Setups

In this section, we describe the main experimental setups. We provide further details in Appendix A.

4.1 Datasets

Since this is the first work on retrieval that utilizes the relational database for augmenting query representations, we design seven different tasks based on the Stack Exchange and Amazon Product Catalog databases available from Fey et al. (2023), and also the H&M database from Robinson et al. (2024).

Stack Exchange This dataset is collected from discussions in Stack Exchange[1], an online website for question-answering, and organized into a relational database consisting of seven tables (such as posts, users, and votes). For this dataset, we design two retrieval tasks, as follows: **Answer Retrieval (Any Answer)** involves retrieving any answer posts

[1] https://stackexchange.com/

made by any users in response to a question post. **Best Answer Retrieval (Best Answer)** is a more challenging task that aims to retrieve a single answer post that has been selected by the owner of the question post. Also, we further consider two different scenarios by dividing the entire dataset by users (**SplitByUser**) or timestamps (**SplitByTime**). For the first setting, training, validation, and test sets are divided by users (there are no overlapping users). Similarly, the later setting splits the dataset according to the timestamp that the post was made. For each retrieval instance, the information before the post timestamp is used to augment the query.

Amazon Product Catalog This dataset is collected from book reviews on the Amazon Product Catalog, which consists of three tables (users, products, and reviews) over the relational database. For this dataset, we introduce **Future Purchase Retrieval (Future Purchase)** as the task, which aims to predict any future book purchases of customers based on their current reviews as well as their previous purchases and reviews. Also, we construct two different settings, namely **ReviewToProduct** and **ProductToProduct**, where the first one uses the review text as a query while the latter one uses the product description as a query for retrieval.

H&M This dataset includes customer and product data across H&M's online shopping platforms, consisting of three tables (articles, customers, and transactions). Similar to the Amazon Product Catalog, we consider the **Future Purchase Retrieval (Future Purchase)** under the **ProductToProduct** setting, whose goal is to predict future product purchases based on the product history as a query.

4.2 Models

We explain the backbone retrieval models and the query augmentation baselines that we compare.

Retrieval Models We use three dense retrievers: **DPR** is a dense retrieval model trained with a pair of a query and its relevant document (Karpukhin et al., 2020); **Contriever** is another dense retriever, but is trained in an unsupervised fashion (Izacard et al., 2022); **BGE-M3** is a recent dense retriever designed to enhance generalization across diverse retrieval tasks (Chen et al., 2024a). As an indicator, we report the results of the sparse retriever (**BM25**).

Augmentation Models We compare our DAQu against relevant query augmentation baselines: **1) No Expansion (No Expan.):** This model uses the query for retrieval without expansion. **2) Query Expansion w/ LLM (Expan. w/ LLM):** This model utilizes the capability of LLMs, prompting them to generate query-related pseudo-documents that are expanded to queries (Wang et al., 2023b). **3) Query Expansion w/ LameR (Expan. w/ LameR):** This model similarly utilizes LLMs but further augments them with query-relevant documents via retrieval for query expansion (Shen et al., 2024). **4) Query Expansion w/ Query associated Table (Expan. w/ Query):** This model expands queries with the information sourced from the query-related single data store (table), following Zhang et al. (2020). **5) Query Expansion w/ User associated Table (Expan. w/ User):** Similarly, this model expands queries with the user-related table, following Buss et al. (2023). **6) Full Metadata Expansion (Expan. w/ Full):** This model concatenates queries with all textual terms of the associated metadata from the database (spanning multiple tables). **7) Query Expansion w/ Retriever (Expan. w/ Retriever):** Similar to Deng et al. (2021), this model also appends the metadata terms to the queries. Yet, before expansion, it employs a retriever (BM25) to select terms that are most relevant to the query, and only these selected terms are appended. **8) DAQu (Ours):** This is our model that augments the query representation by incorporating the metadata representation on a latent space, obtained by graph-structured set-encoding.

4.3 Evaluation Metrics

We use the following metrics: **1) Accuracy@K (Acc@K)** determines the fraction of queries for which the top-k results include at least one relevant document. **2) Recall@K** calculates the percentage of all relevant documents that are present within the top-k results. **3) Mean Reciprocal Rank (MRR)** computes the average of the inverse of the ranks at which the first relevant document is found across queries. **4) Mean Average Precision (MAP)** measures the mean precision score calculated after each relevant document is retrieved, across all queries.

4.4 Implementation Details

We train all retrieval models with a learning rate of 2e-5 with AdamW (Loshchilov and Hutter, 2019). Also, we set λ as 0.7 chosen based on a search within the range of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and randomly sample 30 features for the no-gradient metadata features in our efficient training strategy and 3 features for gradient updates. Regarding metrics, for answer post retrieval on Stack Exchange, which aligns more closely with conventional doc-

Table 1: Results on seven different retrieval scenarios using Stack Exchange, Amazon Product Catalog, and H&M databases.

		StackExchange (Any Answer)				StackExchange (Best Answer)				Amazon (Future Purchase)				H&M (Future Purchase)	
		SplitByUser		SplitByTime		SplitByUser		SplitByTime		ReviewToProduct		ProductToProduct		ProductToProduct	
	Method	Recall@10	Acc@100	Recall@10	Acc@100	MRR	Acc@100	MRR	Acc@100	Acc@500	Recall@1000	Acc@500	Recall@1000	Acc@50	Recall@100
	BM25-Anserini	11.45	28.33	15.79	32.64	9.64	29.49	11.68	34.79	5.71	3.51	15.09	7.48	10.10	3.12
DPR	No Expan.	36.15 ± 0.05	68.09 ± 0.14	35.46 ± 0.55	64.48 ± 0.30	20.87 ± 0.29	56.11 ± 0.09	22.87 ± 0.33	58.25 ± 0.15	6.37 ± 0.49	2.74 ± 0.20	15.54 ± 0.94	7.77 ± 0.24	13.80 ± 1.17	5.52 ± 0.66
	Expan. w/ LLM	32.48 ± 0.26	63.79 ± 0.19	31.66 ± 0.36	60.45 ± 0.43	18.37 ± 0.54	51.60 ± 0.42	20.28 ± 0.32	53.61 ± 0.22	6.37 ± 0.29	2.68 ± 0.10	14.32 ± 0.36	7.67 ± 0.26	13.30 ± 0.29	5.29 ± 0.11
	Expan. w/ LameR	33.77 ± 0.16	65.14 ± 0.30	34.09 ± 1.01	62.34 ± 0.60	20.01 ± 0.38	53.24 ± 0.49	21.60 ± 0.35	55.44 ± 0.76	6.31 ± 0.21	2.62 ± 0.04	15.92 ± 0.57	7.87 ± 0.06	13.97 ± 0.58	6.05 ± 0.04
	Expan. w/ Query	36.70 ± 0.30	69.15 ± 0.22	36.53 ± 0.51	66.60 ± 0.38	20.48 ± 0.38	57.01 ± 0.72	22.57 ± 0.23	58.94 ± 0.41	5.98 ± 0.39	2.58 ± 0.11	16.61 ± 0.29	8.48 ± 0.12	13.64 ± 0.87	5.84 ± 0.28
	Expan. w/ User	36.53 ± 0.06	68.26 ± 0.17	35.65 ± 0.28	65.07 ± 0.15	21.66 ± 0.15	56.74 ± 0.14	23.18 ± 0.06	58.81 ± 0.21	3.48 ± 0.22	2.03 ± 0.10	8.75 ± 0.57	4.68 ± 0.25	13.47 ± 0.58	5.51 ± 0.40
	Expan. w/ Full	38.76 ± 0.21	70.67 ± 0.21	38.75 ± 0.48	67.37 ± 0.45	20.03 ± 0.38	55.00 ± 0.31	21.88 ± 0.14	56.66 ± 0.33	11.04 ± 0.34	**6.10** ± 0.24	14.67 ± 1.21	7.66 ± 0.27	6.57 ± 3.50	1.64 ± 0.55
	Expan. w/ Retriever	38.47 ± 0.34	70.37 ± 0.25	37.83 ± 0.26	66.70 ± 0.15	19.54 ± 0.18	54.08 ± 0.12	21.47 ± 0.26	56.14 ± 0.21	12.56 ± 0.36	5.89 ± 0.25	17.29 ± 0.42	8.42 ± 0.34	9.43 ± 0.58	4.06 ± 0.15
	DAQu (Ours)	**41.80** ± 0.27	**74.11** ± 0.24	**41.67** ± 0.39	**71.72** ± 0.33	**22.05** ± 0.24	**57.81** ± 0.80	**23.70** ± 0.18	**59.24** ± 0.46	**13.07** ± 0.19	5.97 ± 0.27	**17.86** ± 0.39	**9.15** ± 0.10	**15.49** ± 0.29	**6.63** ± 0.15
Contriever	No Expan.	42.08 ± 0.28	73.21 ± 0.15	41.93 ± 0.07	70.08 ± 0.45	25.85 ± 0.15	64.16 ± 0.34	28.37 ± 0.08	64.95 ± 0.15	8.21 ± 0.32	4.63 ± 0.20	17.80 ± 0.45	9.27 ± 0.06	15.15 ± 0.00	5.95 ± 0.00
	Expan. w/ LLM	38.35 ± 0.63	69.35 ± 0.59	38.66 ± 0.29	66.39 ± 0.20	23.27 ± 0.06	59.03 ± 0.12	25.05 ± 0.33	60.32 ± 0.22	8.60 ± 0.31	4.58 ± 0.20	16.82 ± 0.74	9.18 ± 0.24	15.40 ± 0.36	6.20 ± 0.34
	Expan. w/ LameR	38.82 ± 0.04	69.68 ± 0.02	38.78 ± 0.40	67.03 ± 0.03	24.56 ± 0.22	60.12 ± 0.21	25.23 ± 0.18	59.26 ± 0.46	7.26 ± 0.41	3.95 ± 0.24	16.79 ± 0.46	8.73 ± 0.04	15.15 ± 0.00	5.91 ± 0.08
	Expan. w/ Query	41.84 ± 0.31	73.96 ± 0.11	42.92 ± 0.13	71.54 ± 0.45	24.11 ± 0.53	63.39 ± 0.35	27.67 ± 0.11	65.03 ± 0.40	8.93 ± 0.36	4.68 ± 0.17	18.13 ± 0.58	9.31 ± 0.07	**15.66** ± 0.00	6.02 ± 0.06
	Expan. w/ User	42.21 ± 0.36	73.45 ± 0.21	42.26 ± 0.41	70.22 ± 0.20	25.93 ± 0.15	62.87 ± 0.25	28.20 ± 0.12	64.67 ± 0.26	6.34 ± 0.26	2.55 ± 0.15	7.23 ± 0.54	4.35 ± 0.44	11.28 ± 0.29	4.70 ± 0.39
	Expan. w/ Full	45.25 ± 0.24	76.20 ± 0.17	44.43 ± 0.13	72.50 ± 0.18	26.01 ± 0.27	63.59 ± 0.23	28.21 ± 0.10	64.06 ± 0.36	17.23 ± 0.46	8.86 ± 0.22	17.02 ± 0.89	9.37 ± 0.53	5.39 ± 0.29	1.92 ± 0.30
	Expan. w/ Retriever	44.69 ± 0.25	75.52 ± 0.23	44.66 ± 0.27	72.24 ± 0.39	24.71 ± 0.18	62.15 ± 0.24	27.28 ± 0.25	63.52 ± 0.55	17.71 ± 0.22	7.18 ± 0.55	17.71 ± 0.22	9.40 ± 0.21	13.13 ± 0.87	4.99 ± 0.05
	DAQu (Ours)	**49.74** ± 0.26	**80.27** ± 0.23	**50.28** ± 0.49	**78.06** ± 0.38	**26.47** ± 0.26	**65.16** ± 0.33	**28.82** ± 0.07	**65.47** ± 0.58	**18.75** ± 0.91	**9.86** ± 0.46	**19.87** ± 0.44	**10.42** ± 0.67	15.40 ± 0.36	**6.25** ± 0.34
BGE-M3	No Expan.	39.83 ± 0.33	71.08 ± 0.06	39.54 ± 0.44	68.02 ± 0.27	22.37± 0.23	58.41 ± 0.39	22.96 ± 0.20	57.24 ± 0.73	7.59 ± 0.15	3.87 ± 0.03	16.10 ± 0.05	8.29 ± 0.18	14.65 ± 0.00	5.59 ± 0.17
	Expan. w/ LLM	37.57 ± 0.20	67.24 ± 0.47	37.52 ± 0.37	64.29 ± 0.20	19.21 ± 0.13	51.52 ± 0.66	19.95± 0.18	51.72 ± 0.28	8.27 ± 1.60	3.75 ± 0.40	15.98 ± 0.31	8.00 ± 0.09	14.81 ± 0.29	6.05 ± 0.15
	Expan. w/ LameR	38.46 ± 0.34	68.07 ± 0.13	38.06 ± 0.34	65.37 ± 0.19	20.42 ± 0.46	52.41 ± 0.19	21.07 ± 0.16	53.51 ± 0.62	7.32 ± 0.77	3.67 ± 0.66	15.48 ± 0.57	8.20 ± 0.17	14.14 ± 0.00	5.82 ± 0.25
	Expan. w/ Query	39.90 ± 1.16	72.15 ± 0.31	40.64 ± 0.68	70.09 ± 0.26	22.96 ± 0.57	60.32 ± 0.79	23.07 ± 0.50	58.95 ± 0.84	7.41 ± 0.46	3.75 ± 0.36	16.16 ± 0.31	8.25 ± 0.07	15.32 ± 0.29	6.08 ± 0.07
	Expan. w/ User	42.10 ± 0.46	73.13 ± 0.18	41.60 ± 0.23	69.82 ± 0.04	22.84 ± 0.80	59.74 ± 0.93	23.43 ± 0.19	58.47 ± 0.07	4.49 ± 1.19	1.91 ± 0.05	11.79 ± 0.31	5.01 ± 0.27	15.15 ± 0.00	6.01 ± 0.36
	Expan. w/ Full	41.47 ± 0.19	73.00 ± 0.10	41.63 ± 0.90	70.06 ± 0.60	23.42 ± 0.17	58.11 ± 1.06	23.17 ± 0.09	57.29 ± 0.08	13.10 ± 0.05	**7.36** ± 0.47	15.03 ± 1.60	8.12 ± 1.87	4.88 ± 0.29	1.68 ± 0.36
	Expan. w/ Retriever	41.77 ± 0.46	72.76 ± 0.24	41.79 ± 0.23	70.00 ± 0.23	22.84 ± 0.21	58.36 ± 0.36	22.44 ± 0.42	56.25 ± 0.67	12.92 ± 0.26	6.13 ± 0.15	17.56 ± 0.57	8.56 ± 0.17	13.30 ± 1.17	5.49 ± 0.14
	DAQu (Ours)	**44.92** ± 0.22	**75.67** ± 0.05	**45.26** ± 0.39	**73.61** ± 0.07	**24.47** ± 0.45	**61.55** ± 0.18	**24.20** ± 0.01	**59.26** ± 0.24	**14.67** ± 0.88	6.93 ± 0.85	**18.21** ± 0.15	**9.03** ± 0.33	**15.66** ± 0.00	**6.86** ± 0.05

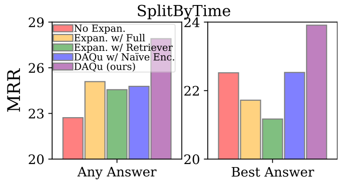

Figure 2: Results of the set encoding strategy of DAQu over naïve encoding, simply aggregating all representations.

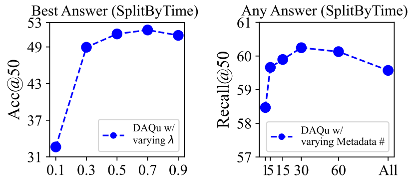

Figure 3: Results by varying lambda values (Left) and the number of metadata features within each category for training (Right).

Table 2: Reranking results following retrieval with DAQu on the SplitByUser scenario of StackExchange (Any Answer).

	Methods	Recall@5	Recall@10	Acc@5	Acc@10
Contriever	**No Expan.**	28.95	36.79	51.00	59.67
	No Expan. + Rerank	35.78	43.09	60.33	67.67
	DAQu (Ours)	35.53	43.78	59.33	68.33
	DAQu (Ours) + Rerank	**42.62**	**49.69**	**69.33**	**75.67**
BGE-M3	**No Expan.**	27.20	34.39	50.33	58.67
	No Expan. + Rerank	34.40	38.99	58.33	62.33
	DAQu (Ours)	32.89	40.40	54.67	63.33
	DAQu (Ours) + Rerank	**40.19**	**45.64**	**62.33**	**66.33**

ument retrieval tasks, we use a diverse range of K values, including 10, 20, 50, and 100. In contrast, for product retrieval with Amazon Product Catalog, where the goal is not only to identify items of interest but specifically those the user will purchase, considering the long-tail nature of product recommendations, we use larger K values of 500 and 1000, following prior work on product retrieval (Li et al., 2021; Wang et al., 2023a; Li et al., 2024). Lastly, we report the average of three different runs.

5 Experimental Results and Analyses

We now present the results and detailed analyses.

Main Results We report the overall results across seven different tasks with multiple databases in Table 1. From this, we find that DAQu outperforms all baselines substantially, demonstrating the effectiveness of our approach that augments queries with their corresponding metadata representations (obtained from graph-based set-encoding). We provide the results with additional metrics in Appendix B.1.

To be specific, our findings reveal that expanding queries with LLMs themselves is suboptimal as their parametric knowledge lacks information specific to each user and its query, which relies instead on general patterns stored within them. In contrast, expanding queries with information from a single source of external data stores (Expan. w/ Query and Expan. w/ User) achieves decent performance improvements over the no-expansion baseline, highlighting the importance of incorporating query-specific and user-specific information during query augmentation. Furthermore, leveraging multiple relational tables from the database, such as Expan. w/ Full and Expan. w/ Retriever, further enhances retrieval performances, which underscores the value of considering interrelated information over the relational database for query expansion.

Notably, the proposed DAQu demonstrates substantial improvements across all tasks over all baselines, highlighting the effectiveness of our proposed set-encoding strategy for incorporating metadata into query representations. For example, in the Answer Retrieval task with Stack Exchange, DAQu achieves performance improvements of 18.73% and 16.91% on SplitByUser and SplitByTime settings, respectively, in Recall@10. Also, DAQu consistently shows superior performance on the Best Answer Retrieval task, which is more complicated (since the model should retrieve the single post that the user would select as the best one, requiring both the query-specific and user-specific information), where diverse expansion models even degrade the performance over the baseline without expansion. Finally, the superior performance of DAQu on the Future Purchase Retrieval task further confirms that it can be applicable to diverse retrieval tasks.

Effectiveness of Set-Encoding To see the effectiveness of the graph-based set-encoding strategy when incorporating the metadata information into the query, we compare it with two types of baselines: appending their textual terms into the query or encoding them without considering the graph

Table 3: Ablation studies involving the removal or addition of each metadata category on Any Answer (SplitByTime), where Q. and A. refer to question and answer posts, respectively.

	Recall		Accuracy	
Metadata Category	**R@20**	**Increase.**	**Acc@20**	**Increase.**
DAQu (Ours)	49.93		54.44	
w/o Comments in Q.	46.75	-6.38%	51.14	-6.06%
w/o Comments in A.	46.06	-7.74%	50.57	-7.11%
w/o Tags in Q.	49.61	-0.63%	54.29	-0.28%
No Expan.	42.22		46.39	
w/ Comments in Q.	45.24	+7.14%	49.69	+7.10%
w/ Comments in A.	47.89	+13.41%	52.31	+12.76%
w/ Tags in Q.	43.60	+3.27%	47.93	+3.31%

Table 4: Results on efficiency, based on elapsed and relative time per query, by varying the number of metadata features for category during inference on Any Answer (SplitByTime).

	Efficiency		Effectiveness	
# of Metadata	**Elpased**	**Relative**	**MAP**	**Acc@100**
No Expan.	0.062	1	22.94	64.15
Expan. w/ Full	0.062	1.002	25.09	67.31
1 per Category	0.073	1.182	24.06	67.99
2 per Category	0.074	1.20	26.69	70.64
3 per Category	0.074	1.205	27.30	71.57
All per Category	0.075	1.218	27.53	71.98

structure. As Figure 2 shows, simply appending the query with additional terms or taking the average of all representations in the metadata without graph structure is not as effective as ours. This demonstrates the efficacy of our two-stage (column- and query-levels) set-based metadata encoding strategy.

Analyses on Metadata Category To investigate how each category of the metadata contributes to overall performance, we conduct ablation studies, reporting the rate of performance increase when excluding or adding each category, with DPR. As Table 3 shows, each category plays a crucial role in performance gains. Also, while each category contributes to improved performance compared to the baseline without expansion, their performances are still not as high as when all categories are used, which implies that the information from each category is complementary to each other. Interestingly, using the 'tags' category (the information within the same table as the query) provides a small improvement, compared to using the 'comments' category from another table, which corroborates our hypothesis that it is important to use knowledge from multiple tables over the relational database.

Analyses on Hyperparameters We explore how varying the lambda value (λ) (balancing the query and metadata representations) impacts the overall results in Figure 3. Specifically, when the lambda value is too low ($\lambda = 0.1$), the model fails to capture the original query's intent. Conversely, a high lambda value ($\lambda = 0.9$) leads to the model overemphasizing the original query over the metadata, thereby underutilizing the meaningful metadata representation, which degrades the performance. Thus, selecting an optimal lambda value is crucial for balancing these aspects to enhance performance.

We further investigate the impact of varying the number of no-gradient metadata features for each category on overall performance, when training the DAQu model. Figure 3 shows that a low count of metadata features per category results in reduced performance, indicating the importance of sufficient features for enhanced results. However, using all metadata features is not only inefficient but also degrades performance. Therefore, it is essential to select the appropriate number of metadata features to optimize model efficiency and effectiveness.

Effectiveness of Reranking with DAQu To see whether DAQu provides high-quality candidate sets for reranking, we conduct an auxiliary analysis, applying the reranker (Sun et al., 2023) on top of the retrieval results from DAQu. As shown in Table 2, reranking leads to substantial performance improvements across all models (while it introduces a slight efficiency trade-off), with DAQu combined with reranking achieving the best results.

Analyses on Inference Efficiency We extend our investigation to the efficiency in inference, by varying the number of metadata features used for query augmentation. As Table 4 shows, although using all the metadata features during inference is effective, it requires more time compared to the model without expansion. By contrast, employing a small number of metadata features enhances efficiency while sacrificing performance. The results indicate that, at a certain point (e.g., 3 features per category), there is a region where we can achieve reasonable performance alongside improved efficiency.

Case Study Lastly, we provide qualitative case studies of our DAQu and its error analysis in Appendix B.7 and Appendix B.8, respectively.

6 Conclusion

In this work, we presented a novel query augmentation framework, DAQu, which enhances the representation of the query with its relevant information within multiple tables over the database. To utilize the metadata features at scale with order invariance, we proposed graph-based set-encoding, which hierarchically aggregates column-level and query-level information. We validated our DAQu on seven different retrieval tasks designed with various databases, showcasing the effectiveness of our database-augmented query representation.

Limitations

While our DAQu framework effectively represents the diverse pieces of query-related metadata (over the relational database) through a graph-structured set-encoding strategy, the process of encoding and aggregating metadata representations at both the column- and query-levels may pose efficiency challenges in real-world applications. To address these concerns, we conducted a detailed analysis of the trade-off between the effectiveness and efficiency of DAQu in Table 4, and showcased that our approach can significantly enhance the effectiveness only with a marginal compensation of the efficiency. On the other hand, this finding still suggests that investigating more advanced methods to further increase run-time efficiency (such as data pruning) would be a valuable direction for future work. Furthermore, the database-augmented retrieval tasks that we designed seem to be quite challenging for the retrieval models. While DAQu generally shows significantly improved performance, there is still a large room for further improving retrieval performance (which we slightly addressed by introducing the reranker in Table 2). Lastly, we wanted to make sure that our framework is validated in realistic retrieval scenarios with real-world large-scale relational databases; however, many such databases are commonly used in enterprise settings and are rarely made publicly available, making it challenging to establish such experimental benchmarking setups. While we validated ours on recently released, real-world relational databases from Stack Exchange, Amazon, and H&M, developing and releasing more databases would be of interest to the community.

We then would like to discuss some interesting avenues for future work, which is orthogonal to the focus of our work and lies far beyond its scopes. First, while our focus is on retrieval, a promising avenue for future research is to extend DAQu to downstream applications, such as Retrieval-Augmented Generation (RAG) (Christmann and Weikum, 2024; Lee et al., 2024), by leveraging the fine-grained and up-to-date user and content information stored in relational databases. Also, it would be valuable to explore a broader challenge faced by many retrieval systems: the trade-off between relevance and exploration. Our work primarily focuses on improving retrieval relevance by leveraging query-associated metadata, as reflected in the performance improvements reported in Table 1. However, in real-world applications, retrieval systems often need to balance relevance with exploration, surfacing diverse or novel content beyond users' historical interests. This challenge, though important, falls outside the scope of our work, as it requires different research assumptions and techniques (such as counterfactual user modeling or diversity-oriented prompting) that are orthogonal to our metadata-driven augmentation framework, leaving for future investigation.

Ethics Statement

A retrieval system can enhance the factual grounding of recent LLMs when it is integrated with them, which helps prevent the generation of plausible but incorrect answers. We believe that, following this line of directions, our DAQu can play a crucial role in diverse retrieval-augmented generation applications. Yet, it is important to note that as relational databases contain substantial amounts of knowledge, including personal information, some potential privacy concerns must be carefully managed when utilizing this information. In other words, further development of filtering strategies that tag and mask personal information across multiple tables before delivery to users or integration with LLMs would be required for real-world applications.

References

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: A survey. *Inf. Process. Manag.*, 56(5):1698–1735.

Jagdev Bhogal, Andrew MacFarlane, and Peter W. H. Smith. 2007. A review of ontology based query expansion. *Inf. Process. Manag.*, 43(4):866–886.

Christopher Buss, Jasmin Mosavi, Mikhail Tokarev, Arash Termehchy, David Maier, and Stefan Lee. 2023. Generating data augmentation queries using large language models. In *Joint Proceedings of Workshops at the 49th International Conference on Very Large Data Bases (VLDB 2023), Vancouver, Canada, August 28 - September 1, 2023*, volume 3462 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1):1:1–1:50.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Peter Baile Chen, Yi Zhang, and Dan Roth. 2024b. Is table retrieval a solved problem? join-aware multi-table retrieval. *arXiv preprint arXiv:2404.09889*.

Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong C. Park. 2023. Improving zero-shot reader by reducing distractions from irrelevant documents in open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3145–3157. Association for Computational Linguistics.

Philipp Christmann and Gerhard Weikum. 2024. Rag-based question answering over heterogeneous data and text. *IEEE Data Eng. Bull.*, 48(4):71–86.

Yang Deng, Yaliang Li, Wenxuan Zhang, Bolin Ding, and Wai Lam. 2021. Toward personalized answer generation in e-commerce via multi-perspective preference modeling. *ACM Transactions on Information Systems (TOIS)*, 40:1 – 28.

Kaustubh D. Dhole and Eugene Agichtein. 2024. Genqrensemble: Zero-shot LLM ensemble prompting for generative query reformulation. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III*, volume 14610 of *Lecture Notes in Computer Science*, pages 326–335. Springer.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. *arXiv preprint arXiv:2405.06211*.

Massimo Esposito, Emanuele Damiano, Aniello Minutolo, Giuseppe De Pietro, and Hamido Fujita. 2020. Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Inf. Sci.*, 514:88–105.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665.

Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. 2023. Relational deep learning: Graph representation learning on relational databases. *arXiv preprint arXiv:2312.04615*, abs/2312.04615.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 4996–5002. ijcai.org.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 512–519. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *Preprint*, arXiv:2305.03653.

Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004a. Umass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004b. Umass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2022. Augmenting document representations for dense retrieval with interpolation and perturbation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 442–452. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, ChaeHun Park, and Jong Park. 2021. Unsupervised document expansion for information retrieval with stochastic text generation.

In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 7–17, Online. Association for Computational Linguistics.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. 2023. Open-wikitable : Dataset for open domain question answering with complex reasoning over table. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8285–8297. Association for Computational Linguistics.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 120–127. ACM.

Meng-Chieh Lee, Qi Zhu, Costas Mavromatis, Zhen Han, Soji Adeshina, Vassilis N. Ioannidis, Huzefa Rangwala, and Christos Faloutsos. 2024. Hybgrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. *arXiv preprint arXiv:2412.16311*, abs/2412.16311.

Millicent Li, Tongfei Chen, Benjamin Van Durme, and Patrick Xia. 2025. Multi-field adaptive retrieval. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Mingming Li, Huimu Wang, Zuxu Chen, Guangtao Nie, Yiming Qiu, Binbin Wang, Guoyu Tang, Lin Liu, and Jingwei Zhuo. 2024. Generative retrieval with preference optimization for e-commerce search. *arXiv preprint arXiv:2407.19829*, abs/2407.19829.

Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based product retrieval in taobao search. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3181–3189. ACM.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023a. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6385–6400. Association for Computational Linguistics.

Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adrià de Gispert, and Gonzalo Iglesias. 2023b. LI-RAGE: late interaction retrieval augmented generation with explicit signals for open-domain table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1557–1566. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yuanhua Lv and ChengXiang Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 1895–1898. ACM.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4089–4100. Association for Computational Linguistics.

George A. Miller. 1992. WORDNET: a lexical database for english. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, USA, February 23-26, 1992*. Morgan Kaufmann.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. Multitabqa: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6322–6334. Association for Computational Linguistics.

Feifei Pan, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, and Peter Fox. 2021. CLTR: an end-to-end, transformer-based system for cell-level table retrieval and table question answering. In *Proceedings of* *the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021 - System Demonstrations, Online, August 1-6, 2021*, pages 202–209. Association for Computational Linguistics.

Feifei Pan, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, and James A. Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *arXiv preprint arXiv:2203.16714*, abs/2203.16714.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan Eric Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. 2024. Relbench: A benchmark for deep learning on relational databases. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 9248–9274. Association for Computational Linguistics.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. 2024. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In *Annual Meeting of the Association for Computational Linguistics*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *ArXiv*, abs/2304.09542.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Binbin Wang, Mingming Li, Zhixiong Zeng, Jingwei Zhuo, Songlin Wang, Sulong Xu, Bo Long, and Weipeng Yan. 2023a. Learning multi-stage multi-grained semantic embeddings for e-commerce search. In *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 411–415. ACM.

Liang Wang, Nan Yang, and Furu Wei. 2023b. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

Zhiruo Wang, Zhengbao Jiang, Eric Nyberg, and Graham Neubig. 2022. Table retrieval may not necessitate table-specific model design. In *Proceedings of the Workshop on Structured and Unstructured Knowledge Integration (SUKI)*, pages 36–46, Seattle, USA. Association for Computational Linguistics.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A. Rossi, Haoliang Wang, and Julian J. McAuley. 2024. Knowledge-aware query expansion with large language models for textual and relational retrieval. *arXiv preprint arXiv:2410.13765*, abs/2410.13765.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. 2017. Deep sets. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3391–3401.

Jiuling Zhang, Beixing Deng, and Xing Li. 2009. Concept based query expansion using wordnet. In *2009 International e-Conference on Advanced Science and Technology*, pages 52–55.

Wenxuan Zhang, Yang Deng, and Wai Lam. 2020. Answer ranking for product-related questions via multiple semantic relations modeling. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 569–578. ACM.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: contextualized query expansion for document re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4718–4728. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*, abs/2101.00774.

Table 5: Data statistics for each task designed with StackExchange, Amazon Product Catalog, and H&M databases.

Task	Setting	Training	Valid	Test
StackExchange				
Any Answer	SplitByUser	128,981	17,132	15,583
	SplitByTime	130,398	15,861	15,437
Best Answer	SplitByUser	43,889	6,106	5,252
	SplitByTime	42,900	6,018	6,329
Amazon Product Catalog				
Future Purchase	ReviewToProduct ProductToProduct	65,797	4,561	5,956
H&M				
Future Purchase	ProductToProduct	24,479	1,133	1,124

A Implementation Details

A.1 Datasets

In this subsection, we provide the additional details for seven tasks (that we design) based on the StackExchange, Amazon Product Catalog, and H&M databases. We first report the detailed statistics of overall datasets in Table 5. Also, in Table 12, we present more fine-grained statistics of each category (column) of the metadata, used for each query. Notably, in this table, we breakdown the metadata features into two categories: ‘total query’ (that includes all the queries in the task) and ‘non-empty query’ (that contains queries with at least one item for each specific metadata category). Lastly, for the schema of each of our considered databases (such as Stack Exchange, Amazon Product Catalog, and H&M), please refer to Figures 4, 5, and 6.

Stack Exchange Recall that, for this database, we design two tasks: **Answer Retrieval (Any Answer)** and **Best Answer Retrieval (Best Answer)**. In this paragraph, we describe which specific metadata categories used for query augmentation. At first, for the Answer Retrieval task, we utilize metadata from the post and comment tables. Specifically, we focus on the tags associated with the current question post and the comments on both the current question and the answer posts. For the Best Answer Retrieval task, we utilize metadata from the post, comment, vote, and user tables. The reason why we utilize more categories for this task is because this task is closely related to the personalized retrieval task (for the user who issues the question post); therefore, we focus on constructing the user-specific metadata. Specifically, we use the total comments made by the user, the ‘aboutme’ information of the user, written question and answer posts, and the voted answer posts by the user. Additionally, we include tags from both the current question post and previously asked question posts. For both tasks, we split the queries with their corresponding metadata into training, validation, and test sets, using a corpus of 3,281,834 documents that contain all posts, according to two different settings. In the SplitByUser setting, we randomly sample users in an 8:1:1 ratio from those who have posted questions with answers provided by others. On the other hand, for the SplitByTime setting, we split the datasets based on the creation timestamp of the question posts. Specifically, we create a training set with question posts written before 2019-01-01, a validation set with posts written after 2019-01-01 but before 2020-01-01, and a test set with posts written after 2020-01-01.

Amazon Product Catalog For this database, we design the **Future Purchase Retrieval (Future Purchase)** task, where we utilize all the user, product, and review tables. Furthermore, we consider the book reviews written from 2013-01-01 to 2016-01-01 (due to the size of the entire corpus), constructing a document corpus using each product’s description, Specifically, we use reviews written in 2013 for the training set, reviews in 2014 for the validation set, and reviews in 2015 for the test set. We then group the reviews written by each customer and randomly sample the customers (since the data before sampling is still very large), selecting 5,000 for the training set, 500 for the validation set, and 500 for the test set. Among two different settings for this task, in the ReviewToProduct setting, each review text (input) is paired with future products (target) that the customer will purchase. For this setting, we incorporate metadata from the previous review text from the review table, and the category, title, and description of both the current and previous products from the product table. In the ProductToProduct setting, we pair the product description of the current review with future products that the customer will buy. We utilize metadata from both the current and previous review texts from the user’s review table, along with the category and title of both current and previous products, and the description of the previous products.

H&M Similar to the Amazon Product Catalog setting, our goal is to predict the future products a customer will purchase by leveraging metadata from both current and previous products, utilizing information from all the user, article, and transaction tables. To achieve this, we consider purchases made between 2020-01-01 and 2020-09-14, using

Table 6: Additional results on seven retrieval tasks with Stack Exchange, Amazon Product Catalog, and H&M databases.

		StackExchange (Any Answer)				StackExchange (Best Answer)				Amazon (Future Purchase)				H&M (Future Purchase)	
		SplitByUser		SplitByTime		SplitByUser		SplitByTime		ReviewToProduct		ProductToProduct		ProductToProduct	
	Method	MAP	MRR	MAP	MRR	Acc@10	Acc@50	Acc@10	Acc@50	Acc@1000	Recall@500	Acc@1000	Recall@500	Acc@100	Recall@50
	BM25-Anserini	7.10	8.61	9.99	11.01	15.50	24.58	18.55	29.14	7.77	2.78	18.39	6.53	12.63	2.49
DPR	No Expan.	23.56 ± 0.03	27.86 ± 0.08	22.72 ± 0.22	25.22 ± 0.24	32.75 ± 0.23	48.63 ± 0.20	35.11 ± 0.60	50.96 ± 0.55	9.23 ± 0.19	1.78 ± 0.27	19.73 ± 0.85	5.98 ± 0.44	14.14 ± 0.88	5.47 ± 0.62
	Expan. w/ LLM	20.97 ± 0.25	24.88 ± 0.30	20.12 ± 0.45	22.45 ± 0.51	28.94 ± 0.85	44.05 ± 0.70	31.31 ± 0.51	46.44 ± 0.30	9.35 ± 0.44	1.67 ± 0.24	19.05 ± 0.22	6.05 ± 0.20	13.30 ± 0.29	5.12 ± 0.04
	Expan. w/ LameR	22.05 ± 0.32	26.13 ± 0.35	22.11 ± 0.65	24.56 ± 0.69	30.59 ± 0.45	45.15 ± 0.76	32.87 ± 0.47	48.29 ± 0.50	9.05 ± 0.21	1.80 ± 0.03	20.57 ± 0.21	6.17 ± 0.40	**15.66** ± 0.00	5.50 ± 0.01
	Expan. w/ Query	23.76 ± 0.07	28.14 ± 0.09	23.67 ± 0.50	26.21 ± 0.51	32.39 ± 0.47	48.74 ± 0.57	35.31 ± 0.24	51.65 ± 0.37	8.57 ± 0.50	1.83 ± 0.29	21.79 ± 0.21	6.59 ± 0.07	14.31 ± 0.29	5.50 ± 0.57
	Expan. w/ User	23.95 ± 0.20	28.14 ± 0.21	22.98 ± 0.10	25.53 ± 0.12	33.57 ± 0.14	49.22 ± 0.20	35.50 ± 0.35	51.68 ± 0.25	5.18 ± 0.71	1.14 ± 0.11	11.25 ± 0.79	3.36 ± 0.25	14.48 ± 0.58	5.03 ± 0.38
	Expan. w/ Full	25.63 ± 0.03	30.15 ± 0.07	25.16 ± 0.11	27.85 ± 0.14	31.44 ± 0.47	47.13 ± 0.41	33.81 ± 0.33	49.27 ± 0.27	16.10 ± 0.92	**4.55** ± 0.24	20.74 ± 1.13	5.54 ± 0.37	7.41 ± 3.79	1.49 ± 0.48
	Expan. w/ Retriever	25.31 ± 0.04	29.79 ± 0.05	24.55 ± 0.05	27.19 ± 0.09	30.98 ± 0.07	46.60 ± 0.31	33.27 ± 0.15	48.72 ± 0.17	17.77 ± 0.36	4.13 ± 0.21	22.65 ± 0.74	6.50 ± 0.13	12.46 ± 1.46	3.13 ± 0.34
	DAQu (Ours)	**27.96** ± 0.23	**32.86** ± 0.10	**27.58** ± 0.31	**30.37** ± 0.35	**33.99** ± 0.25	**50.05** ± 0.33	**36.14** ± 0.42	**52.20** ± 0.47	**18.01** ± 0.29	4.23 ± 0.21	**22.68** ± 1.08	**7.06** ± 0.15	15.49 ± 0.29	**6.62** ± 0.16
Contriever	No Expan.	28.46 ± 0.23	33.23 ± 0.19	28.38 ± 0.28	31.22 ± 0.31	39.71 ± 0.42	56.13 ± 0.33	42.07 ± 0.43	57.90 ± 0.20	12.62 ± 0.73	3.14 ± 0.26	21.76 ± 0.37	7.65 ± 0.19	15.99 ± 0.58	5.69 ± 0.05
	Expan. w/ LLM	25.75 ± 0.70	30.27 ± 0.69	25.83 ± 0.16	28.49 ± 0.15	36.10 ± 0.66	51.42 ± 0.29	37.42 ± 0.61	53.00 ± 0.34	12.68 ± 0.18	3.25 ± 0.23	21.61 ± 0.59	7.17 ± 0.36	16.16 ± 0.00	5.69 ± 0.28
	Expan. w/ LameR	26.14 ± 0.21	30.72 ± 0.14	25.96 ± 0.04	28.74 ± 0.00	37.07 ± 0.22	52.81 ± 0.02	37.50 ± 0.35	52.15 ± 0.70	10.09 ± 0.15	3.07 ± 0.31	21.22 ± 0.21	6.94 ± 0.05	15.32 ± 0.29	5.75 ± 0.01
	Expan. w/ Query	28.15 ± 0.34	32.99 ± 0.41	28.58 ± 0.13	31.43 ± 0.11	37.43 ± 0.26	54.99 ± 0.47	41.11 ± 0.24	57.72 ± 0.14	13.39 ± 0.92	3.29 ± 0.12	22.86 ± 0.29	7.74 ± 0.28	15.91 ± 0.36	5.80 ± 0.12
	Expan. w/ User	28.88 ± 0.21	33.63 ± 0.21	28.07 ± 0.32	30.94 ± 0.29	39.32 ± 0.17	55.92 ± 0.28	42.30 ± 0.42	57.64 ± 0.56	8.57 ± 0.52	1.57 ± 0.23	11.43 ± 0.67	3.16 ± 0.31	12.29 ± 0.29	4.14 ± 0.46
	Expan. w/ Full	31.06 ± 0.16	36.12 ± 0.12	30.12 ± 0.08	33.14 ± 0.08	39.28 ± 0.35	56.04 ± 0.43	41.32 ± 0.15	57.33 ± 0.53	22.65 ± 0.67	7.07 ± 0.14	23.60 ± 0.88	7.14 ± 0.36	6.90 ± 0.58	1.34 ± 0.01
	Expan. w/ Retriever	30.82 ± 0.19	35.76 ± 0.22	30.30 ± 0.32	33.24 ± 0.35	38.09 ± 0.50	54.56 ± 0.25	40.79 ± 0.45	56.42 ± 0.41	22.62 ± 0.22	5.42 ± 0.44	22.62 ± 0.22	7.44 ± 0.04	14.98 ± 0.58	4.29 ± 0.13
	DAQu (Ours)	**35.00** ± 0.33	**40.55** ± 0.41	**34.96** ± 0.53	**38.07** ± 0.57	**40.50** ± 0.16	**57.59** ± 0.58	**42.53** ± 0.06	**58.48** ± 0.51	**25.65** ± 0.44	**7.10** ± 0.29	**25.36** ± 0.50	**8.31** ± 0.23	**17.17** ± 1.43	**5.81** ± 0.49
BGE-M3	No Expan.	26.23 ± 0.49	30.73 ± 0.62	25.72 ± 0.30	28.32 ± 0.29	35.14 ± 0.78	51.30 ± 0.12	35.44 ± 0.22	50.36 ± 0.53	11.52 ± 0.15	2.62 ± 0.06	21.34 ± 0.15	6.61 ± 0.01	14.98 ± 0.58	5.46 ± 0.03
	Expan. w/ LLM	25.14 ± 0.21	29.65 ± 0.19	25.20 ± 0.13	27.89 ± 0.09	30.03 ± 0.30	44.76 ± 0.78	31.18 ± 0.20	45.08 ± 0.36	11.67 ± 1.29	2.50 ± 0.47	20.60 ± 0.36	6.35 ± 0.06	15.15 ± 0.00	5.52 ± 0.17
	Expan. w/ LameR	25.83 ± 0.37	30.29 ± 0.42	25.72 ± 0.20	28.38 ± 0.28	31.31 ± 0.87	45.84 ± 0.37	32.28 ± 0.47	46.41 ± 0.37	10.51 ± 1.13	2.44 ± 0.47	19.49 ± 0.10	6.54 ± 0.23	15.66 ± 0.87	5.16 ± 0.03
	Expan. w/ Query	25.86 ± 0.57	30.25 ± 0.73	26.48 ± 0.41	29.15 ± 0.43	36.39 ± 0.31	52.76 ± 0.89	35.90 ± 0.74	51.93 ± 0.73	11.16 ± 0.46	2.41 ± 0.18	20.60 ± 0.05	6.55 ± 0.15	16.33 ± 0.29	5.62 ± 0.15
	Expan. w/ User	27.41 ± 0.36	31.98 ± 0.38	27.66 ± 0.11	30.41 ± 0.11	36.29 ± 0.96	52.02 ± 1.19	35.91 ± 0.55	51.38 ± 0.54	6.34 ± 1.86	1.33 ± 0.19	15.33 ± 0.10	3.77 ± 0.30	15.99 ± 0.58	5.62 ± 0.03
	Expan. w/ Full	27.35 ± 0.17	32.03 ± 0.16	27.06 ± 0.83	29.78 ± 0.92	35.94 ± 0.27	51.27 ± 1.04	35.46 ± 0.05	50.31 ± 0.30	17.89 ± 0.82	**5.39** ± 0.31	20.98 ± 2.78	5.76 ± 0.61	6.40 ± 0.58	1.36 ± 0.04
	Expan. w/ Retriever	27.91 ± 0.49	32.59 ± 0.44	27.43 ± 0.16	30.14 ± 0.20	35.84 ± 0.02	51.02 ± 0.32	34.22 ± 0.55	49.31 ± 0.89	17.53 ± 0.05	4.29 ± 0.11	23.27 ± 0.36	6.34 ± 0.40	15.99 ± 0.58	4.50 ± 0.28
	DAQu (Ours)	**30.26** ± 0.30	**35.05** ± 0.30	**30.17** ± 0.38	**33.00** ± 0.43	**38.26** ± 1.03	**54.09** ± 0.54	**36.56** ± 0.22	**52.05** ± 0.01	**20.30** ± 1.34	4.78 ± 0.51	**23.36** ± 0.21	**6.86** ± 0.15	**17.51** ± 0.29	**5.81** ± 0.03

Table 7: Metadata statistics (Best Answer, SplitByUser).

Metadata Category	Train (Avg Words per Query)	Test (Avg Words per Query)
Question Posts	2,459.08	1,849.05
Answer Posts	3,690.50	2,934.33
Accepted Answers	1,717.59	1,493.52
Comments	2,844.51	3,169.55
About Me	9.04	10.33
Current Tags	3.06	3.08
Previous Tags	48.36	41.59
Total Words	**10,772.14**	**9,501.45**
Longest Metadata	**307,016**	**439,969**

Table 8: Results for Expan. w/ Full with a special token for each metadata category (DPR, Any Answer, SplitByTime).

Method	Recall@10	Acc@100
No Expan.	35.46	64.48
Expan. w/ Full	38.75	67.37
Expan. w/ Full (w/ Special Tokens)	38.31	67.35
DAQu (Ours)	**41.67**	**71.72**

data from 2020-01-01 to 2020-07-01 as the training set, 2020-07-01 to 2020-08-01 as the validation set, and 2020-08-01 to 2020-09-14 as the test set.

A.2 Models

For DPR (Karpukhin et al., 2020), we follow the implementation by Thakur et al. (2021). For Contriever (Izacard et al., 2022), we further train it from its available checkpoint while using the same architecture as DPR. For a fair comparison, we fix the number of epochs across the same retrieval models for each task and report the average of the three different runs for every model. We use A100 GPU clusters for conducting experiments.

B Experimental Results

B.1 Results with Other Metrics

In addition to our main results in Table 1, we provide the results with other retrieval metrics in Table 6. From this, similar to the results in Table 1, we also observe that our DAQu shows remarkable performance improvements in diverse scenarios.

B.2 Metadata Length Challenges

Our graph-based set-encoding strategy is particularly beneficial when dealing with concatenated textual metadata that may be very long for the encoder to handle. As shown in the metadata statistics in Table 7, the concatenated metadata often results in substantial word counts, with some cases exceeding the token limits of commonly used LLMs, making them impractical for direct processing. Moreover, even when token limits are not exceeded, processing such long contexts can lead to significant computational overhead. These challenges further emphasize the advantages of our graph-based set-encoding approach, which efficiently encodes metadata while preserving its structure and hierarchy.

B.3 Metadata Expansion with Special Token

To evaluate the impact of using special tokens for differentiating metadata categories on retrieval performance for the Full Metadata Expansion baseline (which concatenates a given query with all metadata terms), we extend it by including special tokens for metadata differentiation. As shown in Table 8, the inclusion of special tokens has minimal effect on performance, with Full Metadata Expansion achieving comparable retrieval results regardless of their use.

B.4 Results with Consistent Metrics

In addition to reporting results with diverse metrics to demonstrate the effectiveness of the proposed method across various evaluation criteria, we also provide the results in Table 9 using the same metrics as in Table 3. As shown in Table 9, these results

Table 9: Results with Recall@20 and Acc@20, for Table 3.

	Method	Recall@20	Acc@20
	BM25-Anserini	14.43	17.43
DPR	**No Expan.**	43.09 ± 0.21	50.35 ± 0.29
	Expan. w/ LLM	39.12 ± 0.33	45.97 ± 0.33
	Expan. w/ Query	44.04 ± 0.33	51.28 ± 0.30
	Expan. w/ User	43.31 ± 0.07	50.49 ± 0.13
	Expan. w/ Full	46.20 ± 0.07	53.66 ± 0.09
	Expan. w/ Retriever	45.70 ± 0.03	53.05 ± 0.05
	DAQu (Ours)	**49.54** ± 0.23	**57.13** ± 0.12
Contriever	**No Expan.**	49.20 ± 0.26	56.79 ± 0.28
	Expan. w/ LLM	45.24 ± 0.67	52.64 ± 0.71
	Expan. w/ Query	49.73 ± 0.38	57.49 ± 0.48
	Expan. w/ User	50.00 ± 0.31	57.45 ± 0.46
	Expan. w/ Full	52.57 ± 0.12	60.26 ± 0.10
	Expan. w/ Retriever	52.23 ± 0.24	59.78 ± 0.25
	DAQu (Ours)	**57.33** ± 0.07	**65.05** ± 0.09
BGE-M3	**No Expan.**	47.02 ± 0.44	54.38 ± 0.47
	Expan. w/ LLM	44.08 ± 0.20	51.43 ± 0.24
	Expan. w/ Query	47.34 ± 1.03	54.83 ± 1.19
	Expan. w/ User	48.68 ± 0.15	56.08 ± 0.12
	Expan. w/ Full	48.83 ± 0.02	56.24 ± 0.02
	Expan. w/ Retriever	49.07 ± 0.49	56.47 ± 0.67
	DAQu (Ours)	**52.33** ± 0.04	**60.00** ± 0.19

are consistent with our previous findings, further confirming that our DAQu framework significantly outperforms the baseline methods.

B.5 Results of LameR with BM25

Since LameR is originally designed for sparse retrieval settings (yet we adopt it with the dense retriever to compare against our DAQu framework tailored for dense retrieval), we further explore the variant of LameR with BM25 in Table 10. From this, we find that although LameR provides some improvements coupled with the BM25 retriever, it still lags significantly behind DAQu, which leverages structured metadata in the latent space. These results indicate the importance of dense retrieval, especially in tasks where understanding nuanced relationships in metadata is crucial.

B.6 Results of Expan. w/ Retriever Variant

For the 'Expan. w/ Retriever' baseline, following Deng et al. (2021), we adopt a BM25 model to select metadata terms most relevant to the query and append only those selected terms. To further examine the impact of the retriever used for metadata selection, we replace BM25 with a dense retriever, Contriever. As shown in Table 11, while both retriever-based expansion methods offer moderate gains over the no-expansion baseline, DAQu consistently and significantly outperforms both of them. This highlights the effectiveness of integrating metadata in the latent space, rather than relying solely on term-level expansion.

Table 10: Comparison between BM25-based LameR and our DAQu (with DPR) on the Any Answer and Best Answer tasks.

Method	MRR	Acc@20	Acc@100
Any Answer, SplitByUser			
BM25	8.61	17.43	28.33
BM25 w/ LameR	10.66	21.30	35.14
DAQu (Ours)	**32.86**	**57.13**	**74.11**
Best Answer, SplitByUser			
BM25	9.64	19.42	29.49
BM25 w/ LameR	11.70	23.53	36.48
DAQu (Ours)	**22.05**	**40.40**	**57.81**

Table 11: Results of using Contriever in the 'Expan. w/ Retriever' baseline (Any Answer, SplitByUser).

Method	Recall@10	Acc@100
No Expan.	42.08	73.21
Expan. w/ Retriever (BM25)	44.69	75.52
Expan. w/ Retriever (Contriever)	44.66	76.12
DAQu (Ours)	**49.74**	**80.27**

B.7 Case Study

We conduct a case study to qualitatively compare the effectiveness of our DAQu against the baseline query augmentation methods, provided in Table 13. The first example from the Any Answer retrieval task with the SplitByTime setting presents retrieval results for a user query: selecting optimal activation and loss functions when training an autoencoder on the MNIST dataset. Notably, the challenge here is several important keywords with query-relevant information, such as BCE and MSE, are missing from the original user query. While the baseline expansion models can include such keywords, which can lead to a higher rank of the relevant document (Full Metadata Expansion), Expansion with Retriever results in a lower rank than even No Expansion, due to the exclusion of another essential term, 'Keras'. In contrast, DAQu achieves the highest rank among all baselines, indicating that our method effectively augments all essential information with the metadata representation, by utilizing diverse helpful information sources in a relational database. Similarly, for the Best Answer retrieval task with the SplitByTime setting, given a query such as when normalization or standardization is appropriate, the best answer post explains such cases in terms of 'transformation methods.' Here, our DAQu, which can incorporate the relevant term ‘log transformation’ from the metadata into the query representation, achieves the highest rank. Finally, for the Future Product retrieval task, a user purchased the book ‘Kindergarten-Grade 3’ for their children. In addition, this user’s metadata includes information on several previous purchases tagged ‘Children’s Books.’ In this example, while the No Expansion baseline effectively retrieves the future product with a higher rank, Full Metadata Expansion and Expansion with Retriever do not perform well, suggesting that augmenting metadata with text level adds noise to the retrieval process. Meanwhile, our DAQu effectively exploits only the valuable information on the latent space, achieving the highest rank among all models.

B.8 Error Case Study

As the datasets used in our experiments are collected from real-world applications (e.g., Amazon or StackExchange), the associated metadata is naturally noisy, incomplete, or sometimes weakly relevant to the current query. Then, to better illustrate how our method behaves under such realistic and noisy conditions, we include an error case study in Table 14. As shown in this example, although the query is about evaluating probabilistic predictions, the metadata includes loosely related or distractive content, such as a vague recommendation link and general comments. As a result, the retriever with full metadata expansion (Expan. w/ Full) ranks the correct document much lower (Rank 24). However, our DAQu framework still ranks the correct answer at Rank 4, demonstrating robustness to noise and the ability to effectively utilize metadata signals. This example highlights that while noisy metadata can introduce challenges, DAQu remains effective by learning to selectively incorporate relevant signals (at the representation level), rather than naively aggregating all available metadata in text.

Table 12: Distribution of the metadata features per query for each metadata category for various retrieval tasks.

		Total Query			Non Empty Query		
Setting	**Metadata Category**	**Training**	**Valid**	**Test**	**Training**	**Valid**	**Test**
	StackExchange - Any Answer						
SplitByUser	comments_in_question	1.96	1.95	1.94	3.35	3.37	3.31
	comments_in_answers	2.31	2.45	2.31	3.96	4.14	3.99
	tags	3.00	3.04	3.01	3.00	3.04	3.01
SplitByTime	comments_in_question	2.03	1.69	1.63	3.38	3.19	3.26
	comments_in_answers	2.43	1.89	2.08	4.09	3.46	3.71
	tags	2.97	3.06	3.23	2.97	3.06	3.23
	StackExchange - Best Answer						
SplitByUser	question_posts	14.52	22.15	12.42	18.18	27.07	15.77
	answer_posts	19.77	24.25	13.47	44.79	55.18	30.74
	accepted_answers	7.41	13.41	6.25	10.91	18.68	9.41
	comments	81.28	122.02	84.92	92.86	137.92	97.46
	aboutme	0.33	0.31	0.33	1.00	1.00	1.00
	current_tags	3.06	2.99	3.08	3.06	2.99	3.08
	previous_tags	48.36	66.99	41.59	48.36	66.99	41.59
SplitByTime	question_posts	6.52	7.04	9.96	10.46	11.25	14.94
	answer_posts	7.82	9.35	11.15	27.47	38.98	42.83
	accepted_answers	3.82	3.67	5.36	7.29	7.21	9.77
	comments	31.09	38.59	49.44	54.32	67.36	81.55
	aboutme	0.34	0.29	0.28	1	1	1
	current_tags	3.02	3.10	3.25	3.02	3.10	3.25
	previous_tags	19.52	21.71	32.33	31.31	34.70	48.52
	Amazon Product Catalog						
ReviewToProduct	previous_review_text	8.22	6.97	15.05	11.22	8.94	17.52
	current_product_category	2.90	2.91	2.86	2.99	3.00	2.99
	current_product_title	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_description	1.00	1.00	1.00	1.00	1.00	1.00
	previous_product_category	23.96	20.34	44.16	33.01	26.39	52.68
	previous_product_category	8.22	6.97	15.05	11.22	8.94	17.52
	previous_product_description	8.22	6.97	15.05	11.22	8.94	17.52
ProductToProduct	previous_review_text	8.22	6.97	15.05	11.22	8.94	17.52
	current_product_category	2.90	2.91	2.86	2.99	3.00	2.99
	current_product_title	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_description	1.00	1.00	1.00	1.00	1.00	1.00
	previous_product_category	23.96	20.34	44.16	33.01	26.39	52.68
	previous_product_category	8.22	6.97	15.05	11.22	8.94	17.52
	previous_product_description	8.22	6.97	15.05	11.22	8.94	17.52
	H&M						
ProductToProduct	customer_age	1.00	1.00	1.00	1.00	1.00	1.00
	customer_fashion_news_frequency	1.00	1.00	1.00	1.00	1.00	1.00
	previous_product_description	18.56	45.04	36.44	20.57	46.77	37.28
	current_product_product_type_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_product_group_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_perceived_colour_master_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_department_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_index_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_index_group_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_section_name	1.00	1.00	1.00	1.00	1.00	1.00
	current_product_garment_group_name	1.00	1.00	1.00	1.00	1.00	1.00

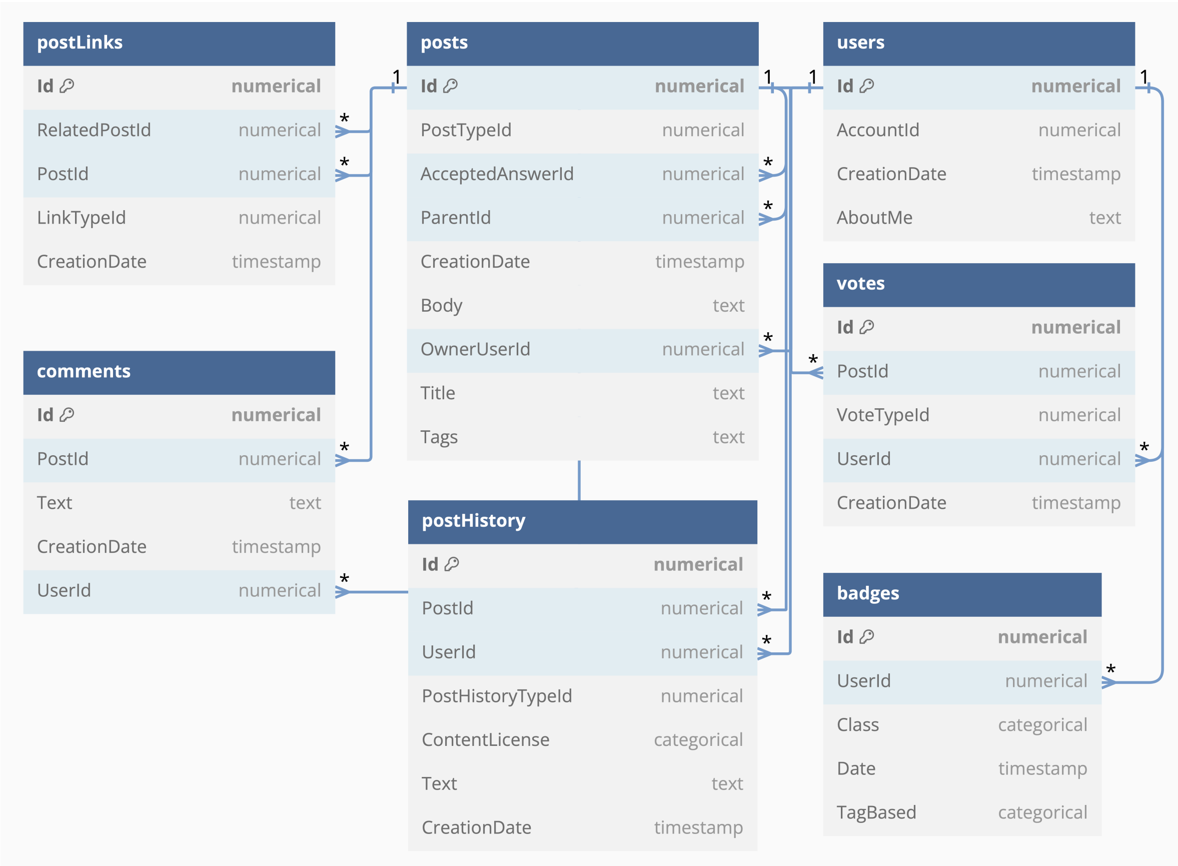

Figure 4: A database schema for Stack Exchange, which is provided from Fey et al. (2023).

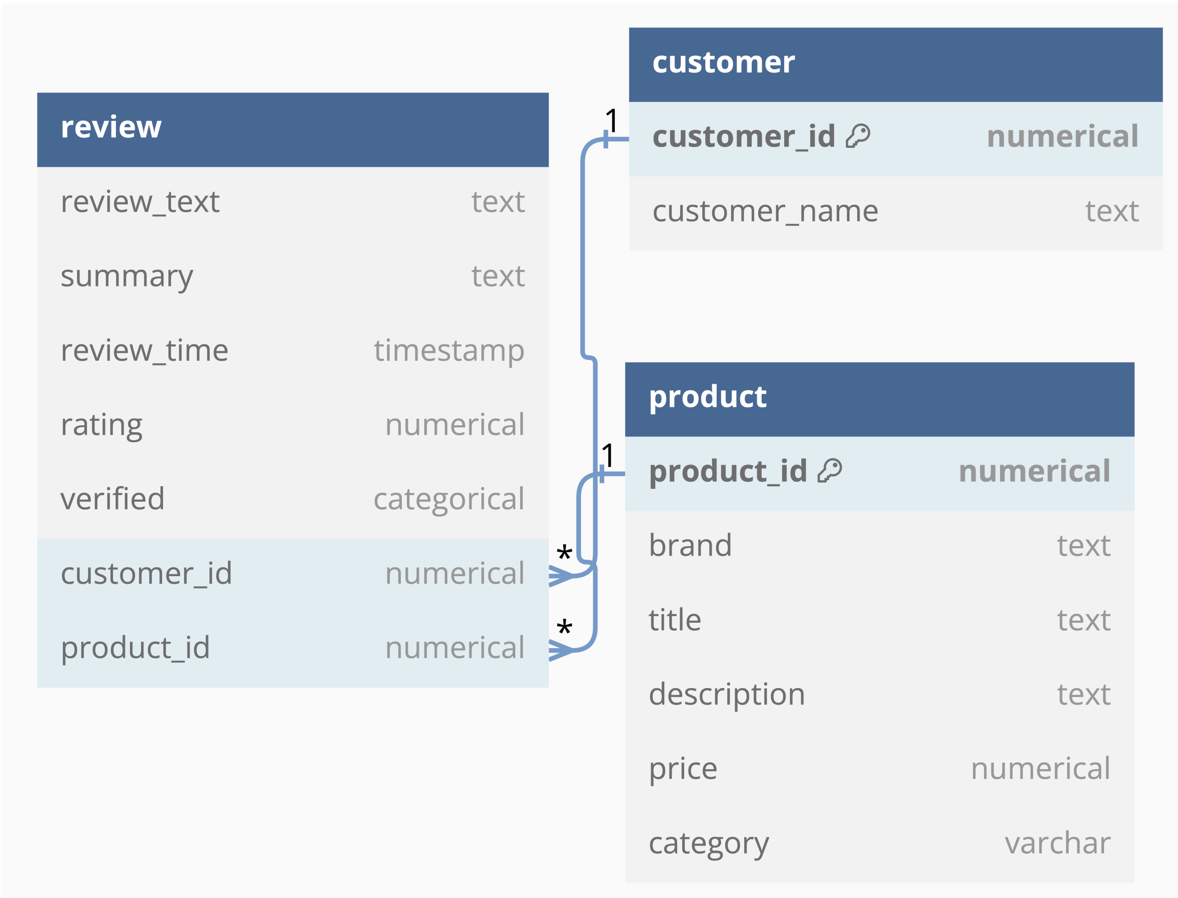

Figure 5: A database schema for Amazon Product Catalog, which is provided from Fey et al. (2023).

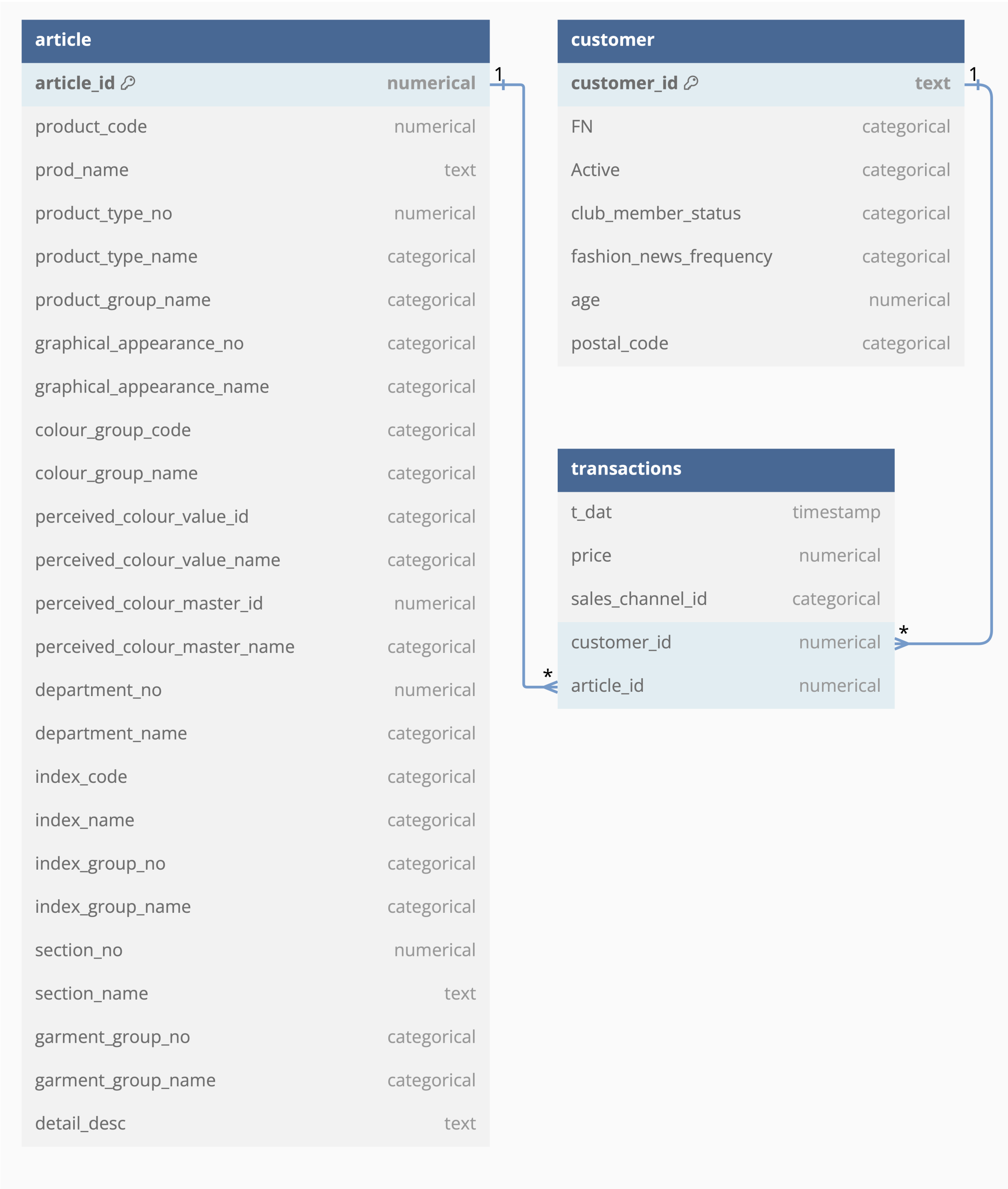

Figure 6: A database schema for H&M, which is provided from Robinson et al. (2024).

Table 13: Case study on three retrieval tasks. In response to the query from the user, notable terms in the Metadata and Answer Post are highlighted in red, which are not in the query but exist only in the metadata and answer posts. Additionally, among those notable terms, some terms that are not covered by the query expansion approach are further highlighted in **bold**.

	StackExchange-Any Answer w/ SplitByTime
Query	**[Title]** Choosing activation and loss functions in autoencoder **[Text]** I am following this keras tutorial to create an autoencoder using the MNIST dataset. Here is the tutorial: <URL>. However, I am confused with the choice of activation and loss for the simple one-layer autoencoder (which is the first example in the link). Is there a specific reason sigmoid activation was used for the decoder part as opposed to something such as relu? I am trying to understand whether this is a choice I can play around with, or if it should indeed be sigmoid, and if so why? Similarily, I understand the loss is taken by comparing each of the original and predicted digits on a pixel-by-pixel level, but I am unsure why the loss is binary crossentropy as opposed to something like mean squared error. I would love clarification on this to help me move forward! Thank you!
MetaData	**[comments in answers by pid]**: ["I wrote about it here, but it was ages ago so I cannot find it now; BCE's properties as a function means it's not the best choice for image data, even in greyscale. Unlike MSE, it is asymmetrically biased against overconfidence, so it systematically underestimates the values, needlessly dimming the output intensities. And, as this question shows, causes unnecessary confusion on top.", "Hmm. I think you may be correct in general, but for this particular use case (an autoencoder), it's been empirically and mathematically shown that training on the BCE and MSE objective both yield the same optimal reconstruction function: <URL> — but that's just a minor detail.", "I cannot load the pdf for some reason, but I'm not surprised - the minima of both losses are the same if your goal is to autoencode a 1:1 match of intensities. It's just not always an optimal loss if your goal is to have a nice-looking image; e.g. MNIST would probably look best with most pixels being either 1 or 0 (in/not in the set of pixels for the character, basically learning a topology)."], **[tags by pid]**: ['neural-networks', 'loss-functions', '**keras**', 'autoencoders']
Answer Post	You are correct that MSE is often used as a loss in these situations. However, the **Keras** tutorial (and actually many guides that work with MNIST datasets) normalizes all image inputs to the range [0, 1]. This occurs on the following two lines: `x_train = x_train.astype(float32) / 255, x_test = x_test.astype(float32) / 255`. Note: as grayscale images, each pixel takes on an intensity between 0 and 255 inclusive. Therefore, BCE loss is an appropriate function to use in this case. Similarly, a sigmoid activation, which squishes the inputs to values between 0 and 1, is also appropriate. You'll notice that under these conditions, when the decoded image is "close" to the encoded image, BCE loss will be small. I found more information about this <URL>.
Retrieval Rank	No Expan. : 26 Expan. w/ Full : 15 Expan. w/ Retriever : 38 DAQu (Ours) : 6
	StackExchange-Best Answer w/ SplitByTime
Query	**[Title]** When to Normalization and Standardization? **[Text]** I see pro-processing with Normalization, which aligns data between 0 and 1, and standardization makes zero mean and unit variance. And multiple standardization techniques follow on.. Any clear definition at what cases what should be used? Thanks in Advance!!
MetaData	**[comments]**: ['hi @onestop, is it ok to take **log transformation** only to skewed columns?'] **[current tags]**:['normalization', 'feature-scaling']
Answer Post	In unsupervised learning, the scaling of the features has a great influence on the result. If a feature has a variance that is many times greater, it can dominate the target function of the algorithm. Therefore, it is of great importance to scale the input data in a way that their variability matches or at least does not contradict the semantics. There are several **transformation methods** to put the features into a comparable form. These use different forms of normalization or standardization according to their context. (...)
Retrieval Rank	No Expan. : 244 Expan. w/ Full : 178 Expan. w/ Retriever : 347 DAQu (Ours) : 105
	Amazon-Future Purchase w/ ProductToProduct
Query	Kindergarten-Grade 3. Fox has composed a simple refrain to celebrate human connections in this lovely picture book. "Little one, whoever you are," she explains, there are children all over the world who may look different, live in different homes and different climates, go to different schools, and speak in different tongues but all children love, smile, laugh, and cry. Their joys, pain, and blood are the same, "whoever they are, wherever they are, all over the world." Staub's oil paintings complement the simple text. She uses bright matte colors for the landscapes and portraits, placing them in gold borders, set with jewels and molded from plaster and wood. These frames enclose the single- and double-page images and echo the rhythm of the written phrases. Within the covers of the book, the artist has created an art gallery that represents in color, shape, and texture, the full range of human experience.
MetaData	**[previous product description]**:["Betsy Snyder's first board book as an author-illustrator, Haiku Baby follows a tiny bluebird, the book's would-be protagonist, as it visits its various animal companions–from an elephant that shades the bird with a parasol to a fox in a meadow and a whale in the ocean. The little bird's story is told primarily in pictures, and through the book's six haiku: rain, flower, sun, leaf, snow, and–of course, it would not be a board book without–the moon, making it ideal for the bedtime line-up. Adorable collage-cut illustrations work nicely with the haiku form to give the book a whimsical, yet serene, feel. And the haiku are light and fun without being too cutesy. Index tabs on the right margin, with pictures that tie to each of the poems (leaf, raindrop, snowflake, etc.), create a unique look, and make it easy for toddlers to flip through the pages on their own without having them stick together like they can with other board books. Snyder excels at visual storytelling and short forms, possibly a talent she honed as a designer/illustrator in the kids' greeting card business. In the world of board books, this slender little volume really stands out"] **[previous product category]**:["Books", "**Children's Books**", "Early Learning"] **[previous review text]**:["My baby loves this book. It has been mouthed, pulled, and thrown many times and still looks new. No tears or running on the pages. No words inside, but has the song on the back incase one does not know it. Can easily make your own story up. My sister washed her book, which you should not do, and it got wrinkled and looks worn down. It did not tear or come apart though", 'Nice little book. Has all the seasons and some weather.']
Future Product	**[Title]** Ten Little Fingers and Ten Little Toes **[Text]** "There was one little baby who was born far away. And another who was born on the very next day. And both of these babies, as everyone knows, had ten little fingers and ten little toes." So opens this nearly perfect picture book. Fox's simple text lists a variety of pairs of babies, all with the refrain listing the requisite number of digits, and finally ending with the narrator's baby, who is 11truly divine" and has fingers, toes, 11and three little kisses/on the tip of its nose." Oxenbury's signature multicultural babies people the pages, gathering together and increasing by twos as each pair is introduced. They are distinctive in dress and personality and appear on primarily white backgrounds. The single misstep appears in the picture of the baby who was "born on the ice." **The child**, who looks to be from Northern Asia or perhaps an Inuit, stands next to a penguin. However, this minor jarring placement does not detract enough from the otherwise ideal marriage of text and artwork to prevent the book from being a first purchase. Whether shared one-on-one or in storytimes, where the large trim size and big, clear images will carry perfectly, this selection is sure to be a hit."
Retrieval Rank	No Expan. : 29 Expan. w/ Full : 162 Expan. w/ Retriever : 765 DAQu (Ours) : 27

Table 14: Error case study for StackExchange (Best Answer).

StackExchange-Best Answer w/ SplitByUser	
Query	**[Title]** How can I determine accuracy of past probability calculations? **[Text]** I do not study statistics but engineering, but this is a statistics question, and I hope you can lead me to what I need to learn to solve this problem. I have this situation where I calculate probabilities of 1000's of things happening in like 30 days. If in 30 days I see what actually happened, how can I test to see how accurately I predicted? These calculations result in probabilities and in actual values (ft). What is the method for doing this? Thanks, CP
MetaData	**[answer posts]**: ['You should check out "area51.stackexchange.com/proposals/117/quantitative-finance?referrer=b3Z9BBygZU6P1xPZSakPmQ2"> area51.stackexchange.com/proposals/117/quantitative-finance?referrer=b3Z9BBygZU6P1xPZSakPmQ2", they are trying to start one on stackexchange.com.'] **[comments]**: ['The Federalist paper is very interesting and describes similar to above answer. Thanks.'] **[current tags]**: ['probability']
Answer Post	What you're looking for are called Scoring Rules, which are ways of evaluating probabilistic forecasts. They were invented in the 1950s by weather forecasters, and there has been some work on them in the statistics community. One thing you could do would be to bin the forecasts by probability range (e.g.: 0–5%, 5%–10%, etc.) and look at how many predicted events in that range occurred. If there are 40 events in the 0–5% range and 20 occur, then you might have problems. If the events are independent, then you could compare these numbers to a binomial distribution. (...)
Retrieval Rank	Expan. w/ Full: 24 DAQu (Ours): 4