# Multi-Facet Blending for Faceted Query-by-Example Retrieval

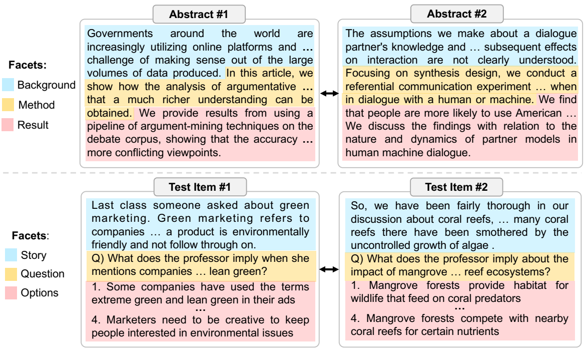
**Heejin Do**[*1], **Sangwon Ryu**[*1], **Jonghwi Kim**[1], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, South Korea
[2]Department of Computer Science and Engineering, POSTECH, South Korea
{heejindo, ryusangwon, jonghwi.kim, gblee}@postech.ac.kr

## Abstract

With the growing demand to fit fine-grained user intents, faceted query-by-example (QBE), which retrieves similar documents conditioned on specific facets, has gained recent attention. However, prior approaches mainly depend on document-level comparisons using basic indicators like citations due to the lack of facet-level relevance datasets; yet, this limits their use to citation-based domains and fails to capture the intricacies of facet constraints. In this paper, we propose a multi-facet blending (FaBle) augmentation method, which exploits modularity by *decomposing* and *recomposing* to explicitly synthesize facet-specific training sets. We automatically decompose documents into facet units and generate (ir)relevant pairs by leveraging LLMs' intrinsic distinguishing capabilities; then, dynamically recomposing the units leads to facet-wise relevance-informed document pairs. Our modularization eliminates the need for pre-defined facet knowledge or labels. Further, to prove the FaBle's efficacy in a new domain beyond citation-based scientific paper retrieval, we release a benchmark dataset for educational exam item QBE. FaBle augmentation on 1K documents remarkably assists training in obtaining facet conditional embeddings.

## 1 Introduction

Query-by-example (QBE), which involves retrieving relevant documents given a query document, is a fundamental technique in both exploratory search (Lissandrini et al., 2019) and recommendation systems (Ostendorff et al., 2020a,b; Lee et al., 2013). However, documents typically include multiple facets distinguished by specific rhetorical units (e.g., *background*, *method*, and *result* of academic paper abstract); thus, querying with the entire document, not identifying the specific facet of interest, can lead to unintentional or irrelevant retrievals (Figure 1). For instance, to recommend exam items similar in question type to a student's incorrect answer, prioritize the *question* facet for retrieval, regardless of *story* or *options*, is required.



Governments around the world are increasingly utilizing online platforms and … challenge of making sense out of the large volumes of data produced. In this article, we show how the analysis of argumentative … that a much richer understanding can be obtained. We provide results from using a pipeline of argument-mining techniques on the debate corpus, showing that the accuracy … more conflicting viewpoints.

The assumptions we make about a dialogue partner's knowledge and … subsequent effects on interaction are not clearly understood. Focusing on synthesis design, we conduct a referential communication experiment … when in dialogue with a human or machine. We find that people are more likely to use American … We discuss the findings with relation to the nature and dynamics of partner models in human machine dialogue.

Last class someone asked about green marketing. Green marketing refers to companies … a product is environmentally friendly and not follow through on.
Q) What does the professor imply when she mentions companies … lean green?
1. Some companies have used the terms extreme green and lean green in their ads
…
4. Marketers need to be creative to keep people interested in environmental issues

So, we have been fairly thorough in our discussion about coral reefs, … many coral reefs there have been smothered by the uncontrolled growth of algae .
Q) What does the professor imply about the impact of mangrove … reef ecosystems?
1. Mangrove forests provide habitat for wildlife that feed on coral predators
…
4. Mangrove forests compete with nearby coral reefs for certain nutrients

Figure 1: Examples of documents with multiple facets.

Accordingly, faceted QBE, which conditions the query document on a specific facet, has garnered recent attention for intent-tailored fine-grained document search (Dunne et al., 2012; Hope et al., 2020; Neves et al., 2019). This task has been predominantly explored in scientific paper retrieval, relying on the vast amount of public corpora where citation labels provide superficial cues (Cohan et al., 2020; Ostendorff et al., 2022; Mysore et al., 2021, 2022). However, those methods are not feasible for other domains (e.g., education or legal), where such citation labels are absent, and large-scale open-source corpora are lacking (Li et al., 2023). Further, the reliance on document-level comparisons often leads to the failure to capture facet constraints, especially for intricate cases (Mysore et al., 2021).

In this paper, we propose a multi-facet blending (FaBle) augmentation method, which dynamically exploits modularity with *decomposing* and *recomposing*. In particular, we first decompose each facet within the document by summary-driven identification, leveraging zero-shot prompting with sLLM. Then, we generate facet-wise *similar* and *dissimilar* facet fragments by self-feeding the decomposed facet summary in recursive prompting.

[*]Equal contribution

*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28577–28590
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Referring to the identified facet guides the synthesis of facet-aware compositions distinguished from other facets. Finally, *recomposition* strategy integrates the synthesized facets to reconstruct facet-conditioned pseudo documents, creating positive–negative pairs for an anchor document. Fable explicitly create facet-specific training sets to assist model training for faceted QBE, eliminating the need for pre-defined facet knowledge or labels.

We target scientific paper abstract retrieval for validation, as it is the sole field providing the benchmark test set for faceted QBE. Aiming to assist in a data-scarce scenario, we employ only 1K documents for augmentation without any citation labels. Experimental results of fine-tuning the SPECTER (Cohan et al., 2020) model with FaBle-augmented pairs are comparable or better to previous models, where more than 1.3M training sets were used for fine-tuning. Notably, FaBle significantly improves the challenging *method* facet, even outperforming the strong prior models. This result highlights that our fine-grained augmentation overcomes the limitations of coarse-grained approaches that ill-capture intricate facets (Mysore et al., 2021).

To further evaluate FaBle's domain scalability and practical efficacy, we present a novel test set for faceted educational exam item retrieval, FEIR, derived from the TOEFL-QA data. Applying FaBle to educational items remarkably improves performance across all facets, demonstrating domain-agnostic effects. We expect FEIR to stimulate future works of faceted QBE in this emerging education domain. Codes and datasets are on GitHub[1].

## 2 Related Work

**QBE** QBE is a fundamental task across diverse fields, such as legal or academic, where document-level findings for recommendation or exploratory search are important (Lissandrini et al., 2019; Ostendorff et al., 2020a,b; Lee et al., 2013). Most prior studies focused on retrieving scientific papers, using large-scale datasets and estimating similarities based on citations (Cohan et al., 2020; Mysore et al., 2021, 2022; Ostendorff et al., 2022). Cohan et al. (2020) introduced the SPECTER to obtain document-level embeddings by measuring similarity via citation graphs, and Ostendorff et al. (2022) used a citation embedding graph combined with neighbor contrastive learning.

**Faceted QBE** Documents typically encompass multiple facets; thus, considering overall document-level relevance may not align with user intent (Do and Lee, 2024). Faceted QBE has emerged to address this, enabling facet-level document comparisons (Neves et al., 2019; Wang et al., 2023). While most studies focus on scientific paper retrieval (Mysore et al., 2021, 2022; Wang et al., 2023), they do not directly train on facet-wise relevance annotated data, as such data is difficult to obtain. Instead, Mysore et al. (2021) utilized an additional 66K citation-based pair for training, and Mysore et al. (2022) used 2.6M co-citation sentences with an auxiliary optimal transport technique. However, relying on abundant domain-specific data and citations restricts its use in low-resource domains.

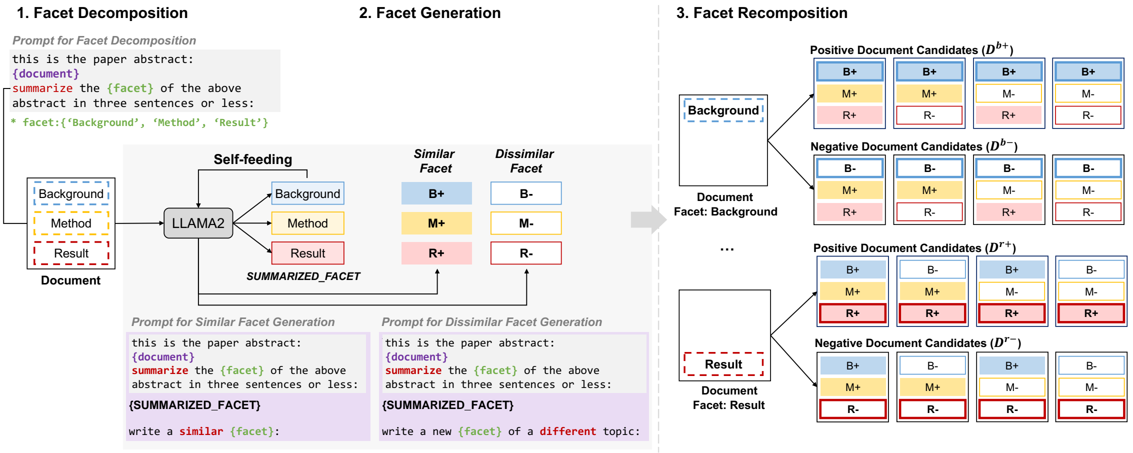
**LLM Augmented Retrieval** LLM-based augmentation techniques have evolved from using GPT-2 (Radford et al., 2019) to GPT-3 (Brown et al., 2020) models to address the lack of relevance annotations. Luu et al. (2021) fine-tune GPT-2 to generate relationships between two scientific papers, assuming in-text citation sentences elucidate their connections. Gao et al. (2023) use GPT-3 to generate hypothetical documents corresponding to desired instructions in a zero-shot manner.

Recently, for faceted QBE, Wang et al. (2023) utilize ChatGPT to annotate the relevance scores of aspect-paper pairs, reducing the burden of human labor. Despite aiming at sub-aspect level similarity evaluation, utilizing ChatGPT for massive datasets still incurs significant costs; thus, they mainly target *testing* faceted QBE, not *training*. Also, as they only contain computer science-related documents, datasets are not generalizable to other fields. Contrarily, by leveraging the capacity of open-source smaller LLM, we eliminate the cost burden and introduce the domain-extendable method.

## 3 FaBle: Multi-facet Blending

For general QBE, obtaining informative representations for query and candidate documents is crucial to effectively retrieve similar documents. To achieve this, model training requires a triplet pair ($D^Q$, $D^+$, $D^-$) comprising a query document, a positive document, and a negative document. In faceted QBE, queries include additional facet conditions; thus, facet-constrained triplet pairs can lead to more precise and focused model training. Unlike prior methods that implicitly construct $D^+$ and $D^-$ based on citations on $D^Q$, we explicitly construct

[1] https://github.com/doheejin/FaBle



Prompt for Facet Decomposition

```
this is the paper abstract:
{document}
summarize the {facet} of the above
abstract in three sentences or less:
* facet:{'Background', 'Method', 'Result'}
```

Prompt for Similar Facet Generation

```
this is the paper abstract:
{document}
summarize the {facet} of the above
abstract in three sentences or less:
{SUMMARIZED_FACET}

write a similar {facet}:
```

Prompt for Dissimilar Facet Generation

```
this is the paper abstract:
{document}
summarize the {facet} of the above
abstract in three sentences or less:
{SUMMARIZED_FACET}

write a new {facet} of a different topic:
```

Figure 2: The overview of the FaBle method and examples of detailed prompts used for scientific paper retrieval.

facet-conditional triplet pairs ($D^{f;Q}$, $D^{f+}$, $D^{f-}$).

FaBle mainly comprises three stages (Fig. 2): decomposition (§3.1), generation (§3.2), and recomposition (§3.3). In this section, we explain examples of scientific paper retrieval, but FaBle is broadly applicable to domains with distinct facets.

### 3.1 Facet Decomposition

To identify each facet, we first decompose the document into multiple facet units. For this, we prompt LLM to summarize a specific facet in a zero-shot manner. We use the publicly available sLLM, LLaMA2-13B (Touvron et al., 2023), taking advantage of open and easy access. By prompting the model to summarize a desired facet within the document, the intended facet-distinct information is extracted. Given a document $D$, summarization prompt $p_{sum}$, and a facet name $f$, where $f \in \{background, method, result\}$, as input, the model generates facet summary $S^f$, which modularize the $f$ facet: $S^f = \texttt{Model}(D, p_{sum}, f)$. Figure 2 describes the detailed prompt, and Figure 3 shows an output summary example. The generated summary highly represents the facet, but it does not mean a *real facet*; instead, it serves as an indicator to guide the subsequent generation stage.

### 3.2 Facet Generation

To generate each facet-specific similar and dissimilar fragment, the same model self-fed the prior prompt used to decompose and its extracted output as shown in Figure 2. Although LLaMA2 has proved proficiency in various generation tasks, its zero-shot performance often lags behind task-specific instruction tuning or GPT-4 (Zhu et al., 2023; OpenAI, 2023). Our self-feeding approach aids in target-oriented generation by referring to the facet-identified summary while eliminating the burden of fine-tuning. In particular, to generate $f$-facet similar component $C^f_{sim}$, the model takes pre-generated summary $S^f$ and the similar-generation prompt $p_{sim}$ as the input. For *dissimilar* component $C^f_{dis}$, the model takes summary $S^f$ and dissimilar-generation prompt $p_{dis}$ as input:

$$C^f_{sim} = \texttt{Model}(D, p_{sum}, f, S^f, p_{sim}) \quad (1)$$

$$C^f_{dis} = \texttt{Model}(D, p_{sum}, f, S^f, p_{dis}) \quad (2)$$

Figure 3 reveals that our two-stage approach results in more target-facet-focused texts (left), while the simple prompting without the facet-identified summary outputs non-target facets mixed in (right).

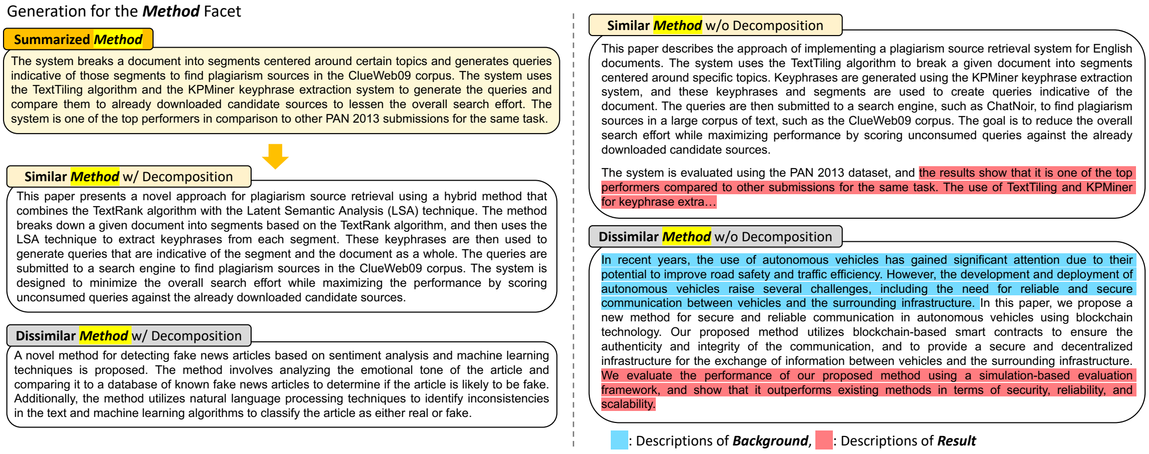
### 3.3 Facet Recomposition

To obtain the negative and positive document pairs for a query document conditioning a specific facet, we combine the generated *similar* and *dissimilar* facet components with a suitable recomposition recipe. The $f$-facet conditional positive $D^{f+}$ and negative $D^{f-}$ documents with total $n$ facets are:

$$D^{f_i+} = [C^{f_1}_{sim|dis}, \ldots, C^{f_{i-1}}_{sim|dis}; \boldsymbol{C^{f_i}_{sim}}; C^{f_{i+1}}_{sim|dis}, \ldots, C^{f_n}_{sim|dis}] \quad (3)$$

$$D^{f_i-} = [C^{f_1}_{sim|dis}, \ldots, C^{f_{i-1}}_{sim|dis}; \boldsymbol{C^{f_i}_{dis}}; C^{f_{i+1}}_{sim|dis}, \ldots, C^{f_n}_{sim|dis}] \quad (4)$$

where [;] denotes concatenation, $F$ is a set of facets, and $f_i \in F - f$ is a facet different from the target facet $f$. Consequently, the triplet pair $(D^{f;Q}, D^{f+}, D^{f-})$ is constructed for the query document $D^{f;Q}$, conditioned on a target facet $f$.



Generation for the ***Method*** Facet

**Summarized *Method***

The system breaks a document into segments centered around certain topics and generates queries indicative of those segments to find plagiarism sources in the ClueWeb09 corpus. The system uses the TextTiling algorithm and the KPMiner keyphrase extraction system to generate the queries and compare them to already downloaded candidate sources to lessen the overall search effort. The system is one of the top performers in comparison to other PAN 2013 submissions for the same task.

**Similar *Method*** w/ Decomposition

This paper presents a novel approach for plagiarism source retrieval using a hybrid method that combines the TextRank algorithm with the Latent Semantic Analysis (LSA) technique. The method breaks down a given document into segments based on the TextRank algorithm, and then uses the LSA technique to extract keyphrases from each segment. These keyphrases are then used to generate queries that are indicative of the segment and the document as a whole. The queries are submitted to a search engine to find plagiarism sources in the ClueWeb09 corpus. The system is designed to minimize the overall search effort while maximizing the performance by scoring unconsumed queries against the already downloaded candidate sources.

**Dissimilar *Method*** w/ Decomposition

A novel method for detecting fake news articles based on sentiment analysis and machine learning techniques is proposed. The method involves analyzing the emotional tone of the article and comparing it to a database of known fake news articles to determine if the article is likely to be fake. Additionally, the method utilizes natural language processing techniques to identify inconsistencies in the text and machine learning algorithms to classify the article as either real or fake.

**Similar *Method*** w/o Decomposition

This paper describes the approach of implementing a plagiarism source retrieval system for English documents. The system uses the TextTiling algorithm to break a given document into segments centered around specific topics. Keyphrases are generated using the KPMiner keyphrase extraction system, and these keyphrases and segments are used to create queries indicative of the document. The queries are then submitted to a search engine, such as ChatNoir, to find plagiarism sources in a large corpus of text, such as the ClueWeb09 corpus. The goal is to reduce the overall search effort while maximizing performance by scoring unconsumed queries against the already downloaded candidate sources.

The system is evaluated using the PAN 2013 dataset, and the results show that it is one of the top performers compared to other submissions for the same task. The use of TextTiling and KPMiner for keyphrase extra…

**Dissimilar *Method*** w/o Decomposition

In recent years, the use of autonomous vehicles has gained significant attention due to their potential to improve road safety and traffic efficiency. However, the development and deployment of autonomous vehicles raise several challenges, including the need for reliable and secure communication between vehicles and the surrounding infrastructure. In this paper, we propose a new method for secure and reliable communication in autonomous vehicles using blockchain technology. Our proposed method utilizes blockchain-based smart contracts to ensure the authenticity and integrity of the communication, and to provide a secure and decentralized infrastructure for the exchange of information between vehicles and the surrounding infrastructure. We evaluate the performance of our proposed method using a simulation-based evaluation framework, and show that it outperforms existing methods in terms of security, reliability, and scalability.

🟦: Descriptions of ***Background***, 🟥: Descriptions of ***Result***

Figure 3: Examples of the generated *similar* and *dissimilar method* facets with our self-fed decomposition (left) and without the (*w/o*) decomposition (right). Directly generating *similar* and *dissimilar* facets without decomposition can lead to the results containing facets other than the intended one, as highlighted.

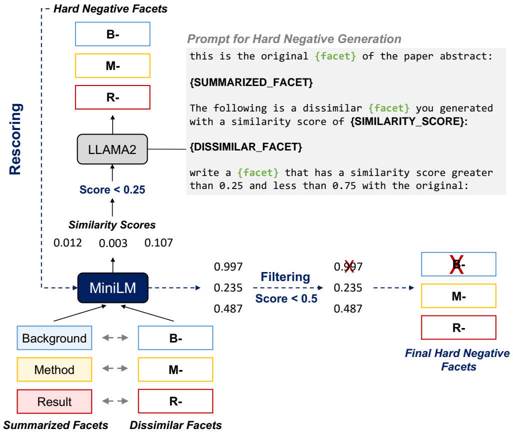
On a single original document with three facets, four $D^{f+}$ and four $D^{f-}$ are generated via facet recomposition. Then, five documents, including the original one, lead to ten $(D^{f;Q}, D^{f+})$ pairs (i.e., five choose two, $\binom{5}{2}$). For each of them, one $D^{f-}$ is selected among four candidates, resulting in a total of forty $(D^{f;Q}, D^{f+}, D^{f-})$ pairs per sample. Note that FaBle operates without any labels, including weak labels like citations or pre-divided facet tags.

### 3.4 Fine-tuning for Faceted QBE

We validate the efficacy of FaBle-augmented triplet pairs in model training via contrastive learning, the widely adopted mechanism for representation learning. Specifically, we employ a pre-trained SciBERT (Beltagy et al., 2019)-based SPECTER (Cohan et al., 2020) to embed the documents. We fine-tune the model with triplet loss to verify whether the synthesized dataset benefits model training. Our loss function $L(D^{f;Q}, D^{f+}, D^{f-})$ is defined as: $max\left\{(\mathrm{d}(D^{f;Q}, D^{f+}) - \mathrm{d}(D^{f;Q}, D^{f-}) + m), 0\right\}$ where d is a distance function, and $m$ is the loss margin hyperparameter. Note that no additional modeling techniques are used to examine the unique effects of the augmentation.

### 3.5 Hard Negative Generation

The significance and efficacy of hard negative mining for retrieval tasks have been widely demonstrated (Xiong et al., 2020; Zhan et al., 2021; Zhang et al., 2021; Zhou et al., 2023). These studies highlight that more challenging negative samples lead to better representation capturing. In this work, we explicitly prompt the LLM to create facets of different topics to generate negative (dissimilar) ones for a specific facet. This may compel the generation of easily distinguishable snippets, potentially leading to the absence of hard negative samples.



```
this is the original {facet} of the paper abstract:

{SUMMARIZED_FACET}

The following is a dissimilar {facet} you generated
with a similarity score of {SIMILARITY_SCORE}:

{DISSIMILAR_FACET}

write a {facet} that has a similarity score greater
than 0.25 and less than 0.75 with the original:
```

Figure 4: Hard negative generation procedure (§ 3.5).

Thus, to enhance the FaBle-generated facets from the perspective of negative sampling, we employ MiniLM[2](Wang et al., 2020), a lightweight cross-encoder model trained on MS MARCO (Bajaj et al., 2016) using knowledge distillation, after Stage 2 (Figure 4). With its proven high performance and efficient inference time (Thakur et al., 2021), MiniLM is ideal for pseudo-relevance scoring. The similarity score, $\texttt{MiniLM}(S^f, C^f_{dis})$, is measured with the summarized facet $S^f$ and the generated *dissimilar* facet component $C^f_{dis}$ inputs. The output score reflects how closely the gener-

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

|  | Orig |  |  | FEIR |  |
|---|---|---|---|---|---|
|  | Train | Valid | Test | Query | Cands |
| Full | 717 | 124 | 122 | - | - |
| Story | 150 | 24 | 24 | 8 | 23 |
| Question | 717 | 124 | 122 | 8 | 80 |
| Options | 717 | 124 | 122 | 8 | 70 |

Table 1: Summary of the original TOEFL-QA dataset (*Orig*) and the FEIR test set. *Story* is a shared facet among multiple question-options sets.

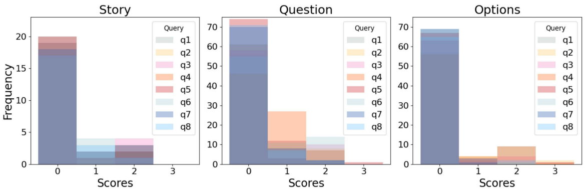


Figure 5: Score label distributions per query by facet.

ated facet fragments align with the original facets. Based on the score distribution, we regard the negative samples with a similarity score below 0.25 as *easy negatives*. Here, we aim to regenerate those samples to have a specific score distribution of 0.25–0.5 for hard negative mining. To control the relevance level, we notify the LLM of the current similarity score by including it in the prompt, inspired by recent studies that incorporate exact numeric values in instructions (Ribeiro et al., 2023; Zhang et al., 2024). We then measure the MiniLM scores for the regenerated facets and identify those below 0.5 as hard negatives. The recomposition process in Stage 3 is applied to the added facets, yielding the final supplemental hard negatives.

## 4 FEIR

The benchmark test set for faceted QBE is absent in domains other than scientific paper retrieval. This gap leads to a shortage of related studies in other fields, such as educational item retrieval, where each item comprises multiple facets. Even when items share similar *Questions*, their *Stories* and *Options* may differ, requiring fine-grained search queries. To validate the scalability of FaBle and support future research, we introduce a Faceted Educational exam Item Retrieval (FEIR) test set for the underexplored language education domain.

**Dataset Construction** We employ exam items from the publicly available TOEFL-QA[3](Chung et al., 2018; Tseng et al., 2016) dataset, a representative English as a Foreign Language (EFL) exam, to build the FEIR. The dataset contains 963 TOEFL listening QA items, and we utilize 122 test set items for constructing the FEIR test set (Table 1). Inspired by CSFCube (Mysore et al., 2021), which has 16 queries per facet, and given our limited original dataset, we form 8 query items for each facet (total 24 queries). To ensure diversity in relevance scores, we evaluated each sample's similarity with MiniLM scores and sequentially selected eight unique samples with the largest standard deviations in their score distributions. Each facet contains four conversation-type and four lecture-type queries. For candidate selection in the *story* facet, where data is limited, we use all 23 remaining items except the query item. In the *question* and *options* facets, we choose 80 and 70 items, prioritizing those with the highest standard deviations after removing the query items.

**Relevance Annotation** To annotate relevance between facet-specific query-candidate pairs, we hired three experts: a language-learning major university professor and two English specialists from Upwork[4]. Each facet was assigned to two different experts. Following detailed guidelines and rating criteria (Appendix A, E), they rated the relevance of each query and candidate item on a 0–3 scale, similar to Mysore et al. (2021); the rounded average of two ratings is the final score. Figure 5 shows the score distribution of candidates per query, with a minority being labeled between 1 and 3. This trend mirrors the CSFCube test set, where an average of 36.9 candidates per query are rated 1, and 9.8 candidates receive scores of 2 or 3. We examine the inter-annotator agreement by measuring the correlations between two annotators' labels: Kendall's $\tau$, Spearman's $\rho$, and Pearson's $r$. The facet-average values are 0.474, 0.492, and 0.557, respectively ($p<0.05$), indicating positive agreements (Chiang and Lee, 2023).

## 5 Experiments

**Data and Settings** We use only 1017 random paper abstracts from the 81.1M papers in the open-source S2ORC[5] corpus (Lo et al., 2020), having metadata, abstracts, and full text of academic papers. However, we do not use any annotated information in this work. By deliberately limiting the initial data to a small amount (approximately 0.00125%), we aim to validate that our method is

[3] https://github.com/iamyuanchung/TOEFL-QA/

[4] https://www.upwork.com/

[5] https://allenai.org/data/s2orc

| CFSCUBE Facets | Background | | Method | | Result | | Aggregated | |
|---|---|---|---|---|---|---|---|---|
| Model | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP |
| SentBERT-PP | 60.80 | - | 33.40 | - | 52.35 | - | 48.57 | - |
| SentBERT-NLI | 54.23 | - | 31.10 | - | 51.30 | - | 45.39 | - |
| CoSentBert | 61.27 | 35.78 | 38.77 | 19.27 | 50.68 | 32.15 | 50.68 | 28.95 |
| SCINCL | 70.02 | 49.64 | 46.61 | 27.14 | 61.70 | 41.83 | 59.24 | 39.37 |
| SPECTER-ID | 69.22 | - | 42.76 | - | 60.40 | - | 57.22 | - |
| TSASPIRE$_{\text{Spec}}$ | 70.22 | 49.58 | 48.20 | 28.86 | 64.39 | 42.92 | 60.71 | 40.26 |
| OTASPIRE$_{\text{Spec}}$ | 71.04 | 50.56 | 46.46 | 27.64 | 67.38 | 44.75 | 61.41 | 40.79 |
| TS+OTASPIRE$_{\text{Spec}}$ | 70.99 | 51.79 | 47.60 | 26.68 | 64.82 | 43.06 | 60.86 | 40.26 |
| SPECTER | 66.70 | **43.95** | 37.41 | 22.44 | 56.67 | 36.79 | 53.28 | 34.23 |
| +FaBle (Ours) | **67.38**<br>±0.28 | 42.66<br>±0.32 | **44.97**<br>±0.16 | **25.98**<br>±1.05 | **58.10**<br>±1.78 | **38.60**<br>±1.31 | **56.60**<br>±0.57 | **35.60**<br>±0.52 |
| SPECTER-COCITE$_{\text{Scib}}$ | 68.71 | 48.40 | 46.79 | 26.95 | 59.68 | **38.93** | 58.16 | 37.90 |
| SPECTER-COCITE$_{\text{Spec}}$ | 70.03 | **49.99** | 45.99 | 25.60 | 59.95 | 37.33 | 58.38 | 37.39 |
| +FaBle$_{\text{Spec}}$ (Ours) | **70.09**<br>±0.09 | 45.93<br>±0.54 | 49.14<br>±0.95 | 30.90<br>±0.89 | 60.88<br>±0.86 | 38.08<br>±0.20 | **59.79**<br>±0.26 | 38.11<br>±0.31 |
| +FaBle$_{\text{Spec}}$+HN (Ours) | 69.48<br>±0.83 | 46.03<br>±0.60 | **49.43**<br>±1.11 | **32.57**<br>±1.32 | **61.09**<br>±0.37 | 38.14<br>±0.64 | 59.76<br>±0.75 | **38.73**<br>±0.66 |

Table 2: Evaluation results on CSFCube test set. *SPECTER-COCITE*$_{\text{Spec}}$ and *SPECTER-COCITE*$_{\text{Scib}}$ are the SPECTER- and SciBERT (Beltagy et al., 2019)-initialized model trained with co-citation dataset, respectively. *+FaBle* and *+FaBle*$_{\text{Spec}}$ denote fine-tuning on the above *SPECTER* and *SPECTER-COCITE*$_{\text{Spec}}$, respectively. *+FaBle*$_{\text{Spec}}$*+HN* is the addition of **H**ard **N**egative samples. **Bold**: the highest among baseline and proposed methods, underline: the highest score in each column, ±: standard deviation of three runs.

| | Background | Method | Result |
|---|---|---|---|
| Summarized Facet ($S^f$) | **0.756** | 0.668 | 0.685 |
| Similar Component ($C^f_{sim}$) | **0.736** | 0.634 | 0.649 |

Table 3: Averaged similarity scores between the entire document and each facet (denoted as $S^f$ and $C^f_{sim}$).

effective in practical data-scarce settings. As the CSFCube comprises scientific papers in the computer science domain, we also select abstracts from the same field. Applying the FaBle with 1K documents, 40 triplet document pairs are generated per facet for a single document, resulting in 40.68K triplet pairs. To apply FaBle for education exam items, we use 717 items from the TOEFL QA training set, creating total 28.68K pairs As the dataset already has facet labels, we directly employ Stages 2 and 3. Detailed settings are in Appendix B.

**Baselines** Most studies on faceted QBE have used or fine-tuned the SPECTER (Cohan et al., 2020) model; hence, we adopt it as our baseline. Our primary aim is to evaluate the efficacy of facet-specific augmentation in data-scarce settings rather than resorting to supplementary methods for fine-grained QBE. Thus, our comparisons focus on the baseline models and those fine-tuned with FaBle-augmented data. We train two versions: the original SPECTER and SPECTER-COCITE$_{\text{SPEC}}$. The latter is similar to SPECTER but was additionally trained on 1.3M co-citation datasets from Mysore et al. (2022) with 2–3 point aggregation across queries. We also assess whether the FaBle-assisted model is comparable to other strong models for faceted QBE, with further details in Appendix C.

**Evaluation** For evaluation, we use CSFCube[6] (Mysore et al., 2021) test set, which provides annotations for faceted QBE on computer science papers. 50 query abstract–facet pairs are assigned relevance scores (0–3). We use the FEIR set to evaluate the educational exam item. For metrics, we employ normalized discounted cumulative gain at rank K (NDCG@K) and mean average precision (MAP). In particular, we report NDCG$_{\%20}$, computing at 20% of the query pool size, following prior works (Wang et al., 2013; Mysore et al., 2021, 2022). For the FEIR with fewer queries and candidates, we also report the NDCG$_{\%10}$.

## 6 Results

Table 2 shows the main results of FaBle across three facets. Incorporating FaBle with SPECTER enhances performance in all facets, yielding notable average gains of 3.4% in NDCG$_{\%20}$ and 1.4% in MAP. For SPECTER-COCITE, fine-tuning the model with FaBle also improves the performance, highlighting our assistance in model training.

**Facet-Specific Results** The *method* facet, widely recognized as the most challenging primarily due to its focus on procedural descriptions of technical concepts, encountered difficulties in assessing similarity with prior models (Mysore et al., 2021, 2022).

[6] https://github.com/iesl/CSFCube

| CFSCUBE Facets | Background | | Method | | Result | | Aggregated | |
|---|---|---|---|---|---|---|---|---|
| Model | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP |
| SPECTER-COCITE$_{\text{Spec}}$ | 70.03 | **49.99** | 45.99 | 25.60 | 59.95 | 37.33 | 58.38 | 37.39 |
| +FaBle$_{\text{Spec}}$ | **70.09** | 45.93 | **49.14** | **30.90** | **60.88** | **38.08** | **59.79** | **38.11** |
| +FaBle-RN$_{\text{Spec}}$ | 69.61 | 46.58 | 46.82 | 28.62 | 59.83 | 37.47 | 58.48 | 37.32 |

Table 4: Ablation study results. While *FaBle* includes the generation of *dissimilar* facets in Stage 2, *FaBle-RN* selects Random facets as Negatives. +*FaBle*$_{\text{Spec}}$ and +*FaBle-RN*$_{\text{Spec}}$ denote fine-tuning on the above model.

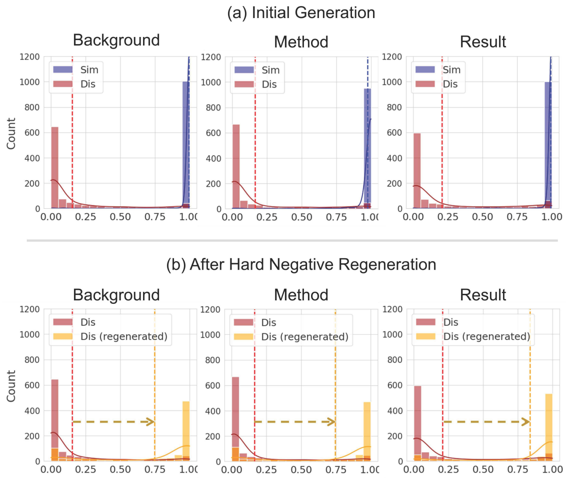


Figure 6: MiniLM score distributions of generated *Sim* and *Dis* facets. (a) Initial distributions; (b) shifts in negative samples after regeneration; dashed line: mean.

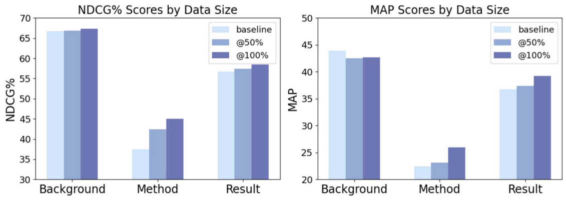


Figure 7: Comparison of NDCG$_{\%20}$ (left) and MAP (right) performances by the dataset size per facet.

In this context, the remarkable enhancements in the *method* facet are noteworthy: an increase of 7.6% in NDCG$_{\%20}$ and 3.5% in MAP scores over SPECTER. Moreover, FaBle with the SPECTER-COCITE achieved a 3.4% rise in NDCG$_{\%20}$ and a 7.0% increase in MAP scores, even outperforming the robust ASPIRE models, trained on ≈32 times greater dataset than FaBle and employ co-citations labels with additional optimal transport techniques. Unlike them, FaBle leverages the knowledge embedded within LLMs trained on massive corpora to make intrinsic judgments about *similarity* by individual facets. This enables the generation of sentences that deliberately mirror or distort procedural domain knowledge, resulting in sophisticated candidate construction, even for complex facets. Thus, the synthesized data can contribute to more discriminative representations for retrieval.

However, the *background*, already achieved high scores (52.27% higher NDCG$_{\%20}$ than the *method* on SPECTER-COCITE$_{\text{Spec}}$), shows modest improvements, with a slight decline in MAP. This outcome is attributable to the comprehensive nature of the paper's background (Andrade, 2011), which existing coarse-grained systems can sufficiently capture; small gaps across all models also support this. Further, Mysore et al. (2022) noted that the stronger correlation between background contents and the paper's overall topic leads to the success of general models that incorporate whole abstract-level representations. We find this dependency by examining similarity scores between each decomposed and generated facet and full document. Table 3 reveals that the *background* facet has a higher similarity to the entire document than other facets, explaining the less pronounced impact of our facet-specific approach on it.

In the *result* facet, the impact of FaBle is evident, although not as large as the *method*. This outcome aligns with prior models' performance, falling between the other two facets. Some *result*s can be easily identified as similar by common phrase overlaps, while others demand a detailed interpretation of the query (Mysore et al., 2021), where our sophisticated processing can be effective. The *result*s are contextually dependent on other facets, as they typically discuss *method*-driven observations or *background*-posed problem-solving. Consequently, their similarities are shaped by overall abstract relevance (Mysore et al., 2022). The superiority of multi-match-based OTASPIRE over single-match-based TSASPIRE in the *result* facet supports this. Thus, enriching FaBle with auxiliary methods addressing global-level similarity can be beneficial.

## 7 Analysis and Discussion

**Impact of Hard Negatives** We investigate the impact of hard-negative generation (§ 3.5). Before analyzing, we examine how our hard-negative sampling altered the score distribution of exist-

|  | Story |  |  | Question |  |  | Options |  |  | Aggregated |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | NDCG$_{\%20}$ | NDCG$_{\%10}$ | MAP | NDCG$_{\%20}$ | NDCG$_{\%10}$ | MAP | NDCG$_{\%20}$ | NDCG$_{\%10}$ | MAP | NDCG$_{\%20}$ | NDCG$_{\%10}$ | MAP |
| SCINCL | **69.15** | 71.88 | 60.79 | 29.64 | 23.05 | 19.00 | **80.26** | **78.81** | 58.67 | 59.68 | 57.91 | 46.15 |
| +FaBle | 69.11 | **76.04** | **61.67** | **29.91** | **26.34** | **23.70** | 80.15 | 78.51 | **58.87** | **59.72** | **60.30** | **48.08** |
|  | ±0.70 | ±0.00 | ±0.03 | ±0.34 | ±0.52 | ±0.01 | ±0.65 | ±0.65 | ±2.70 | ±0.30 | ±0.04 | ±0.90 |
| SPECTER | 61.85 | 65.62 | 64.20 | 27.57 | 21.75 | 17.43 | 78.74 | 78.18 | 53.75 | 56.05 | 55.18 | 45.13 |
| +FaBle | **64.36** | **65.62** | **64.40** | **28.10** | **22.61** | **19.35** | **80.43** | **78.35** | **55.66** | **57.63** | **55.53** | **46.47** |
|  | ±0.63 | ±0.00 | ±0.02 | ±0.10 | ±0.00 | ±0.08 | ±0.00 | ±0.00 | ±0.02 | ±0.19 | ±0.00 | ±0.02 |

Table 5: Evaluation results on FEIR test set. FaBle denotes fine-tuning the SPECTER with our augmented dataset.

|  | **Question** |
|---|---|
| Orig | **Why does the professor** ask the man to **come to her office**? |
| Sim | What would be an appropriate **reason why the professor** might **invite** the student **to her office**? |
| Dis | What are some benefits of studying abroad? |
|  | **Options** |
| Orig | 1.The effect of the decrease in **temperatures** on wetlands<br>2.The use of computer models to analyze **temperature** patterns<br>3.The theory that land development **affected** the **climate** of South Florida<br>4.The importance of the citrus industry to the South Florida economy |
| Sim | 1.The impact of urbanization on local **ecosystems**<br>2.The role of water management practices in shaping **regional climates**<br>3.The **influence** of agricultural activities on atmospheric conditions<br>4.The effects of deforestation on biodiversity and **climate** |
| Dis | 1.The impact of social media on teenagers' self-esteem<br>2.The benefits of meditation for mental health<br>3.The history of the civil rights movement in the United States<br>4.The role of parental involvement in student academic achievement |

Table 6: Generated *Sim* and *Dis* facets of *Question* and *Options*. Relevant terms are highlighted in **bold**.

ing negatives. Figure 6 exhibits that regeneration shifted the average to around 0.75 points, aligning with our goal of acquiring more challenging samples. We only select samples below 0.5 as hard negatives to differentiate from positives. Table 2 (FaBle$_{\text{Spec}}$+HN) indicate that hard negatives for a specific facet, regenerated to have a higher similarity score, remarkably assist *method*-faceted retrieval but not in the others. Creating high-similarity negative samples to a specific facet may hinder the relevance recognition on the general facets like *background* and *result*. Yet, for facets demanding a fine-grained approach, auxiliary optimizing with hard negatives can boost contrastive learning (Qu et al., 2021; Santhanam et al., 2022; Ostendorff et al., 2022; Formal et al., 2022).

**Comparison with Random Sampling** We compared the efficacy of directly generating negative facets to random sampling (Table 4). In particular, we replaced dissimilar facet fragments created for each document with randomly selected original facets from other documents. Original facets are not defined in the document; thus, we utilize the summarized facets from Stage 1. Table 4 indicate that FaBle, which integrates generating *dissimilar* component as specific-facet-tailored negatives, achieves markedly better performance than FaBle-RN, which employs random sampling. Hence, our subtle negative sampling may be a key for faceted QBE, aligning with contemporary research that emphasizes the advantages of strategic negative sampling over random approaches (Qu et al., 2021; Zhan et al., 2021; Zhou et al., 2023).

**Effects of the Data Size** We examine how the amount of augmentation affects model performance. For 50%, we randomly select half the original document (0.5K out of 1K), creating 20K triplet pairs with FaBle. Figure 7 reveals that increasing the data size consistently enhances NDCG%20 and MAP scores. For both metrics, the *Background* facet shows reasonably high scores even at the base level, implying that the model itself could represent this comprehensible facet well; hence, fine-tuning on larger data moderately impacts the model performance. Meanwhile, the *Method* facet, indicated to be underrepresented in the baseline model by exhibiting lagged performance behind the other two facets, shows a clear improving tendency as the amount of FaBle-augmented data increases. Thus, tailoring data size to the specific needs of individual facets is essential for training optimization.

**Results on FEIR** Table 5 presents the experimental results of fine-tuning SPECTER on FaBle-augmented data with educational items. FaBle brings in performance improvements in all facets, with the substantial 3.6% NDCG and 3.4% MAP increases in the *question* facet, mirroring trends of the CSFCube results (Table 3). Generally, when holistically finding similar items using a coarse-grained approach, *question*, which comprises a single sentence, is more likely to be overlooked than options' four sentences and a story of a paragraph constituting multiple sentences. In contrast, FaBle constructs facet-specific positive and negative documents with modularized combinations, allowing even less prominent facets to be targeted. In the qualitative analysis for generation (Table 6), similar components are content-relevant to the original, while dissimilar ones shift to irrelevant topics, implying FaBle's ability to fit intentions.

## 8 Conclusion

We introduce FaBle, a multi-facet blending augmentation that aids in direct model training for faceted QBE. By modularizing facets by decomposing and recomposing, FaBle effectively synthesizes pseudo-documents that match user-intended facets, eliminating the need for pre-set annotations. In particular, our blending method introduces cost-efficient recomposition techniques to puzzle together initially generated facet combinations; thus, it enables diverse augmentation without requiring additional generation. FaBle improves the retrieval performances, particularly in the salient facet, surpassing models trained on much larger datasets. In addition, we release a FEIR test set for the language education domain, demonstrating FaBle's generalizability.

## Limitations

Currently, we assume data scarcity by applying FaBle on a small amount of data to evaluate the assistance in real-world settings where open corpora are limited. However, as we observed the performance-improving trends with the increased amount of datasets, augmenting with more original data could lead to further enhancements. Secondly, in prior faceted QBE works, statistical tests are not provided, which may be attributed to the small test set size (e.g., 16-17 queries per facet in CSFCube), confining statistical power (Morgan, 2017; Serdar et al., 2021). To address this limitation and enhance reproducibility, we conducted experiments with three different seeds and reported average scores with standard deviations. Furthermore, to practically investigate the significance of FaBle, we examined the proportion of queries where performance remained equal or improved. Among the aggregated queries, 70.83%, 70.83%, and 75% showed increased NDCG$_{\%20}$, NDCG$_{\%10}$, and MAP scores over SPECTER, respectively, demonstrating that FaBle is effective for the majority of queries.

## Ethical Statement

We follow the guidelines outlined in the ACL Code of Ethics. Our work did not utilize private datasets, and it does not include any confidential personal information. For the human annotation, we provide fair compensation for two human experts, paying $135 as a fixed price.

## Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-RS-2020-II201789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (50%), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH)) (50%).

## References

Chittaranjan Andrade. 2011. How to write a good abstract for a scientific paper or conference presentation. *Indian journal of psychiatry*, 53(2):172.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *NAACL HLT*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Heejin Do and Gary Geunbae Lee. 2024. Aspect-based semantic textual similarity for educational test items. In *Artificial Intelligence in Education*, pages 344–352, Cham. Springer Nature Switzerland.

Cody Dunne, Ben Shneiderman, Robert Gove, Judith Klavans, and Bonnie Dorr. 2012. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2022. From distillation to hard negative sampling: Making sparse neural ir models more effective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2353–2359.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel Weld, Marti Hearst, and Jevin West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 135–143, Online. Association for Computational Linguistics.

Joonseok Lee, Kisung Lee, and Jennifer G Kim. 2013. Personalized academic research paper recommendation system. *arXiv preprint arXiv:1304.5457*.

Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. Thuir@ coliee 2023: Incorporating structural knowledge into pre-trained language models for legal case retrieval. *arXiv preprint arXiv:2305.06812*.

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-based search: a new frontier for exploratory search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1411–1412.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Kelvin Luu, Xinyi Wu, Rik Koncel-Kedziorski, Kyle Lo, Isabel Cachola, and Noah A. Smith. 2021. Explaining relationships between scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2130–2144, Online. Association for Computational Linguistics.

Charity J Morgan. 2017. Use of proper statistical techniques for research studies with small samples. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, 313(5):L873–L877.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470, Seattle, United States. Association for Computational Linguistics.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. Csfcube-a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, pages 124–135.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020a. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020b. Pairwise multi-class document classification for semantic relations between wikipedia articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 127–136, New York, NY, USA. Association for Computing Machinery.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the*

*2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Leonardo FR Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. *arXiv preprint arXiv:2310.10623*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Ceyhan Ceran Serdar, Murat Cihan, Doğan Yücel, and Muhittin A Serdar. 2021. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, 31(1):27–53.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *INTERSPEECH*.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Doris-mae: Scientific document retrieval using multi-level aspect-based queries. *CoRR*, abs/2310.04678.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yucheng Zhou, Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Guodong Long, Binxing Jiao, and Daxin Jiang. 2023. Towards robust ranker for text retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5387–5401, Toronto, Canada. Association for Computational Linguistics.

Hanlin Zhu, Andrew Cohen, Danqing Wang, Kevin Yang, Xiaomeng Yang, Jiantao Jiao, and Yuandong Tian. 2023. End-to-end story plot generator. *Preprint*, arXiv:2310.08796.

## A Relevance Annotation for FEIR

For FEIR relevance annotation, experts rated the relevance degree of each query and candidate item within the facet on an integer scale from 0 to 3, similar to Mysore et al. (2021), according to the following guidelines:

- 3 (Near Identical): A strong and clear correlation exists between the facet of a query item and a candidate item. Significant overlap in content, background, context, or theme indicates a high association level.
- 2 (Similar): An apparent degree of connection is observed between the facet of a query item and a candidate item. Shared elements or themes suggest a moderate level of association.
- 1 (Related): A superficial connection exists but is minimal. There may be slight thematic or contextual similarities, but the items are mainly independent.

- 0 (None or Irrelevant): Items that do not meet the criteria for the above categories should be labeled as 0.

## B Experimental Settings

For all the experiments, we report the average results conducted by three runs with different seeds, `{22,222,2222}`. The batch size is set as 30, following the settings of the previous model (Mysore et al., 2022). We use a 1e-5 learning rate and two epochs for academic paper retrieval and a 1e-6 learning rate and two epochs for education items. Margin $m$ for the triplet loss is set as 1. Fine-tuning and model inference are performed using an A100-SXM4-40GB GPU and take approximately 2 hours. For LLaMA, we used the LLaMA2-13B chat model[7]. To fine-tune SPECTER, we split the generated dataset into training and validation sets with a 9:1 ratio.

## C Baseline Models

We examine the comparability of our method's performance with competitive models that exhibited strong results in facet-conditional retrieval. In particular, we report the results of a robust faceted QBE model, ASPIRE (Mysore et al., 2022), and its various comparisons, outlined in Mysore et al. (2021) and Mysore et al. (2022). The reported TSASPIRE$_{\text{Spec}}$ is a SPECTER-based single-match method with textual supervision, OTASPIRE$_{\text{Spec}}$ is a multi-match method utilizing optimal transport, and TS+OTASPIRE$_{\text{Spec}}$ combines both approaches, as a multi-task and multi-aspect method. They are trained with 1.3M training sets. Their comparison models, `SentBERT-PP`, `SentBERT-NLI`, and `CoSentBert`, are MPNET-1B[8] based sentence embedding models. SPECTER-ID results are also reported, fine-tuned with 660K in-domain papers that fit the CSFCube test set.

## D Evaluation Metrics

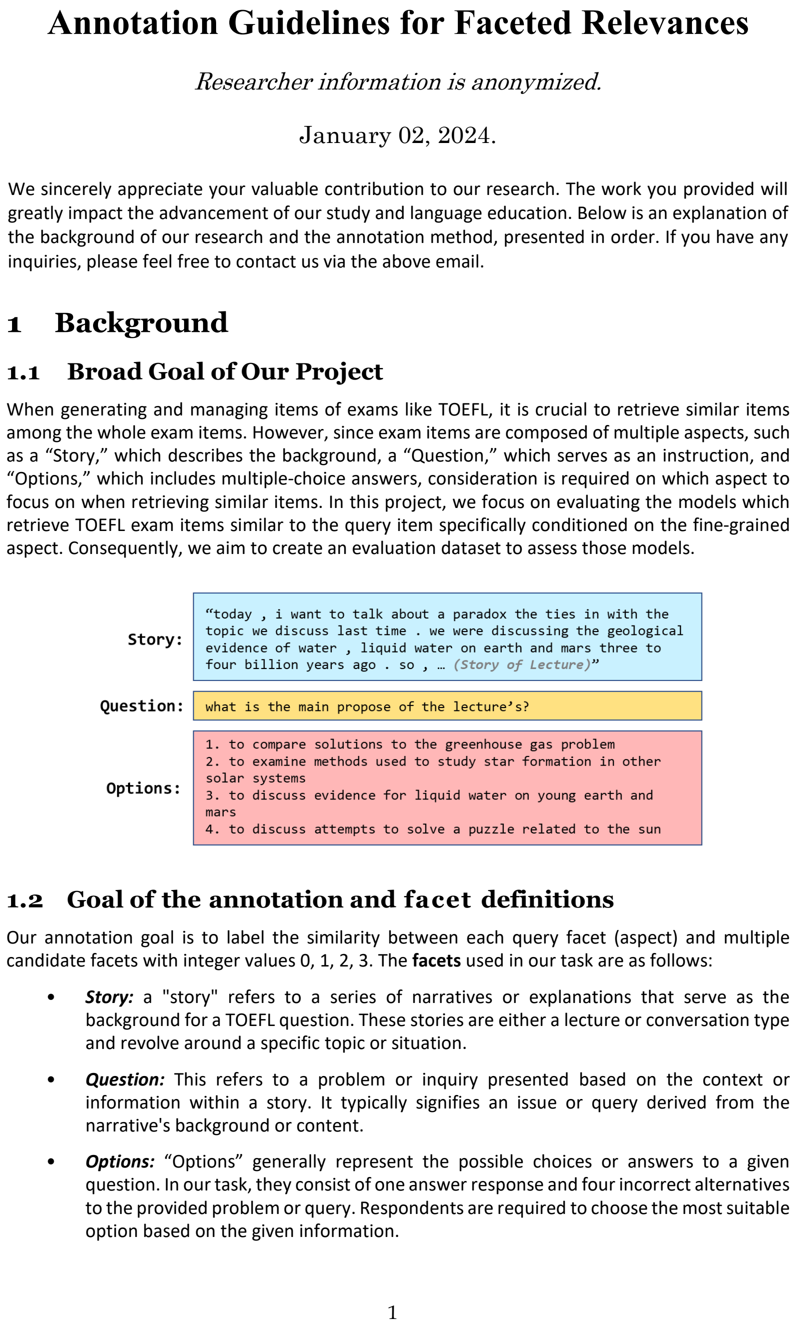
For evaluation, we employ normalized discounted cumulative gain at rank K (NDCG@K) and mean average precision (MAP), well-known retrieval metrics. In particular, we set $K = p * |C|$ where $p \in (0, 1)$ and report NDCG$_{\%20}$, which denotes computing at 20% of the query pool size, following existing research (Wang et al., 2013; Mysore et al., 2021, 2022). The NDCG metric reflects the graded relevance scores of items to assess the ranking quality, offering a more nuanced perspective than binary metrics such as precision or recall, particularly when the dataset is annotated with multiple relevance scores. Given that our test sets, CSFCube and FEIR, have multiple numeric relevance annotations, NDCG would be the most suitable metric. Specifically, for the FEIR test set, which has fewer queries and candidates than CSFCube, we also report the NDCG$_{\%10}$ results, which compute at 10% of the query pool size.

## E Detailed Annotation Guidelines for FEIR

Figures 8 and 9 illustrate the specific guidelines provided to the annotators for FEIR annotation.

[7] https://ai.meta.com/llama/

[8] MPNET-1B is pre-trained over 1B text pairs



# Annotation Guidelines for Faceted Relevances

*Researcher information is anonymized.*

January 02, 2024.

We sincerely appreciate your valuable contribution to our research. The work you provided will greatly impact the advancement of our study and language education. Below is an explanation of the background of our research and the annotation method, presented in order. If you have any inquiries, please feel free to contact us via the above email.

## 1 Background

### 1.1 Broad Goal of Our Project

When generating and managing items of exams like TOEFL, it is crucial to retrieve similar items among the whole exam items. However, since exam items are composed of multiple aspects, such as a “Story,” which describes the background, a “Question,” which serves as an instruction, and “Options,” which includes multiple-choice answers, consideration is required on which aspect to focus on when retrieving similar items. In this project, we focus on evaluating the models which retrieve TOEFL exam items similar to the query item specifically conditioned on the fine-grained aspect. Consequently, we aim to create an evaluation dataset to assess those models.

**Story:** “today , i want to talk about a paradox the ties in with the topic we discuss last time . we were discussing the geological evidence of water , liquid water on earth and mars three to four billion years ago . so , … *(Story of Lecture)*”

**Question:** what is the main propose of the lecture’s?

**Options:**
1. to compare solutions to the greenhouse gas problem
2. to examine methods used to study star formation in other solar systems
3. to discuss evidence for liquid water on young earth and mars
4. to discuss attempts to solve a puzzle related to the sun

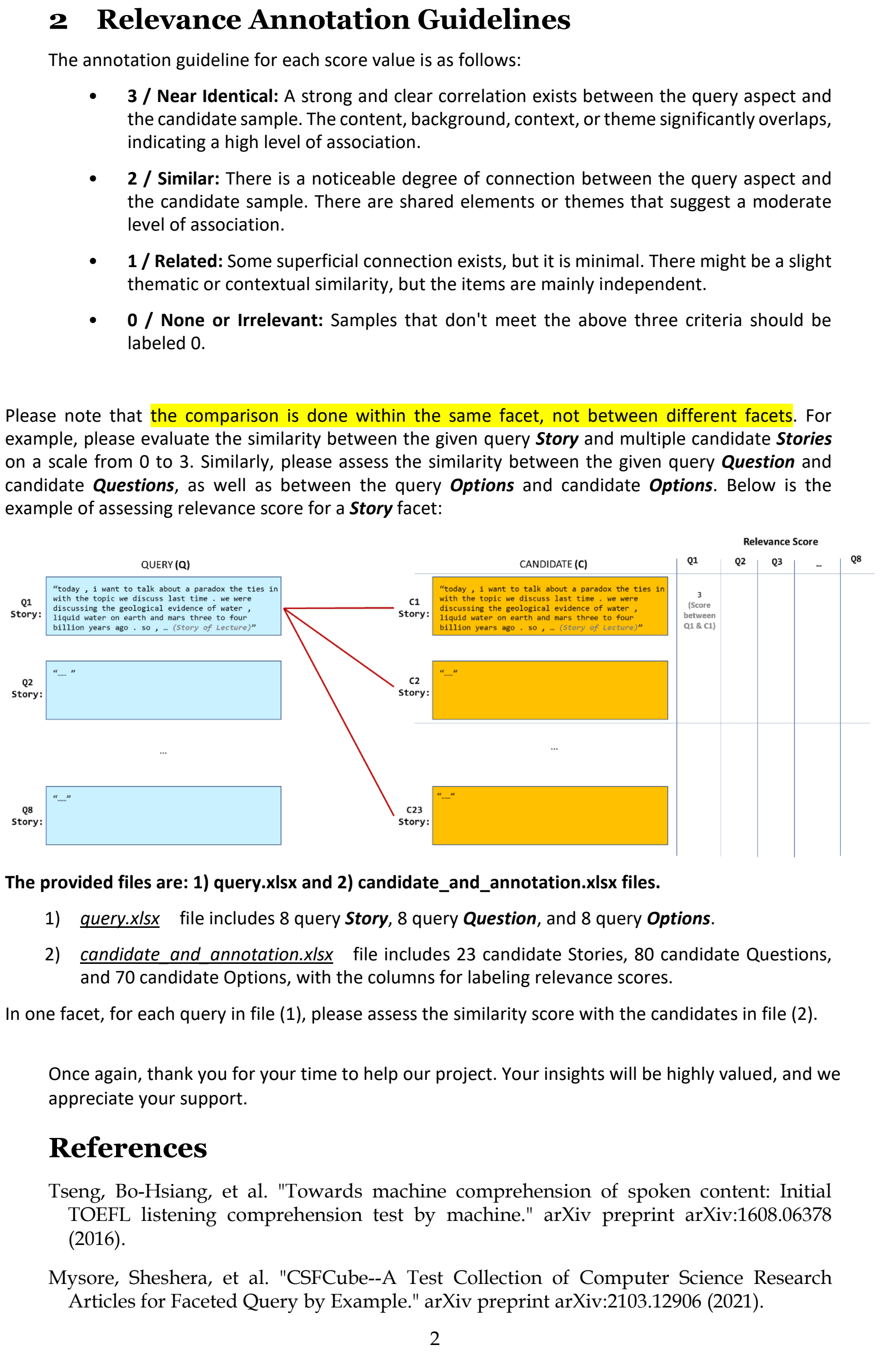
### 1.2 Goal of the annotation and facet definitions

Our annotation goal is to label the similarity between each query facet (aspect) and multiple candidate facets with integer values 0, 1, 2, 3. The **facets** used in our task are as follows:

- ***Story:*** a "story" refers to a series of narratives or explanations that serve as the background for a TOEFL question. These stories are either a lecture or conversation type and revolve around a specific topic or situation.
- ***Question:*** This refers to a problem or inquiry presented based on the context or information within a story. It typically signifies an issue or query derived from the narrative's background or content.
- ***Options:*** “Options” generally represent the possible choices or answers to a given question. In our task, they consist of one answer response and four incorrect alternatives to the provided problem or query. Respondents are required to choose the most suitable option based on the given information.

1

Figure 8: The first page of the provided guidelines for FEIR annotation.



## 2 Relevance Annotation Guidelines

The annotation guideline for each score value is as follows:

- **3 / Near Identical:** A strong and clear correlation exists between the query aspect and the candidate sample. The content, background, context, or theme significantly overlaps, indicating a high level of association.
- **2 / Similar:** There is a noticeable degree of connection between the query aspect and the candidate sample. There are shared elements or themes that suggest a moderate level of association.
- **1 / Related:** Some superficial connection exists, but it is minimal. There might be a slight thematic or contextual similarity, but the items are mainly independent.
- **0 / None or Irrelevant:** Samples that don't meet the above three criteria should be labeled 0.

Please note that the comparison is done within the same facet, not between different facets. For example, please evaluate the similarity between the given query ***Story*** and multiple candidate ***Stories*** on a scale from 0 to 3. Similarly, please assess the similarity between the given query ***Question*** and candidate ***Questions***, as well as between the query ***Options*** and candidate ***Options***. Below is the example of assessing relevance score for a ***Story*** facet:

Relevance Score
QUERY (Q)
CANDIDATE (C)
Q1
Q2
Q3
...
Q8
Q1 Story:
"today , i want to talk about a paradox the ties in with the topic we discuss last time . we were discussing the geological evidence of water , liquid water on earth and mars three to four billion years ago . so , … (Story of Lecture)"
Q2 Story:
"…… "
...
Q8 Story:
"……"
C1 Story:
"today , i want to talk about a paradox the ties in with the topic we discuss last time . we were discussing the geological evidence of water , liquid water on earth and mars three to four billion years ago . so , … (Story of Lecture)"
3 (Score between Q1 & C1)
C2 Story:
"……"
...
C23 Story:
"……"

**The provided files are: 1) query.xlsx and 2) candidate_and_annotation.xlsx files.**

1) *query.xlsx* file includes 8 query ***Story***, 8 query ***Question***, and 8 query ***Options***.
2) *candidate_and_annotation.xlsx* file includes 23 candidate Stories, 80 candidate Questions, and 70 candidate Options, with the columns for labeling relevance scores.

In one facet, for each query in file (1), please assess the similarity score with the candidates in file (2).

Once again, thank you for your time to help our project. Your insights will be highly valued, and we appreciate your support.

## References

Tseng, Bo-Hsiang, et al. "Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine." arXiv preprint arXiv:1608.06378 (2016).

Mysore, Sheshera, et al. "CSFCube--A Test Collection of Computer Science Research Articles for Faceted Query by Example." arXiv preprint arXiv:2103.12906 (2021).

Figure 9: The second page of the provided guidelines for FEIR annotation.