sciDataQA: Scientific Dataset Recommendation for Question Answering

Anonymous ACL submission

Abstract

In order to advance scientific discovery, it is essential to answer scientific questions regarding a particular field of study. However, these questions might not be answered easily with just a few words and might mislead scientists, delaying scientific discovery. In this paper, we propose to recommend scientific datasets instead of directly answering each question. We introduce sciDataQA, a novel scientific dataset recommendation dataset with 43466 scientific datasets and 244128 questions, including each dataset's title, citation information, summary, and abstract. We construct the dataset with large pre-trained language models and utilize a contrastive-learning-based approach to filter the low-quality questions. Based on this dataset, we develop a novel recursive retrieval approach for scientific dataset recommendation. Further, we illustrate how our dataset can be used to study citation prediction and improve existing scientific QA systems. Extensive experiments show the effectiveness of our recursive retrieval approach and the improvement in the low-resource setting of two existing scientific QA systems with our dataset.

1 Introduction

Question answering (QA) has become an increasingly important task due to the massive amount of data from a variety of resources (Wang, 2022). Scientific question-answering systems aim to answer questions about a specific scientific domain and could be critical for scientific discovery (Clark et al., 2018; Mihaylov et al., 2018; Lu et al., 2022). These questions are often answered by performing machine reading comprehension on scientific literature (Khashabi et al., 2020; Xu et al., 2021; Huang et al., 2022). However, in contrast to traditional QA, scientific questions might not be answered easily according to existing literature; many new research problems are never studied in the literature; an incorrect answer might mislead scientists and delay scientific discovery.

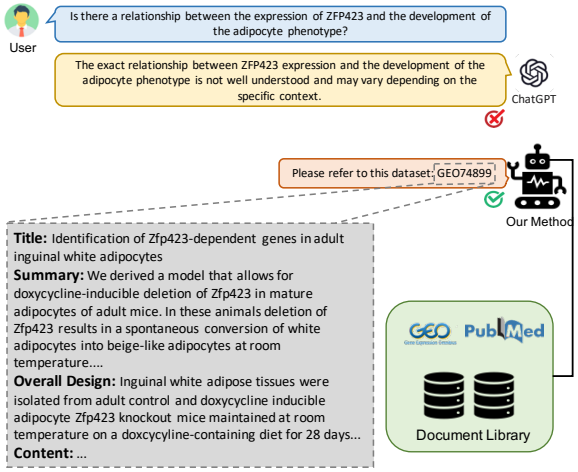

Figure 1: An illustration of the comparison between traditional science QA bot and our dataset recommendation approach.

To circumvent this challenge, we propose to recommend scientific datasets instead of directly answering this question. The input of our scientific QA system is still a scientific question. The output will be a dataset that we recommend scientists analyze in order to answer this question. This dataset recommendation task mimics the scientific discovery process of raising a hypothesis and then retrieving relevant datasets to answer this hypothesis. Compared to answering the question directly, recommending a dataset is more feasible and can offer more flexibility for scientists to analyze it. On the other hand, recommending a dataset requires us to comprehend not only the question but also the scientific dataset.

As this scientific dataset recommendation task has not been systematically studied before, we first construct a large-scale dataset, sciDataQA, which contains 244128 questions and 43466 scientific datasets from Gene Expression Omnibus (GEO) (Edgar et al., 2002). Each question is paired with one dataset from GEO. For each dataset, we first identified the collection of scientific papers that

have used this dataset. We then extracted the mentions of this dataset and used pre-trained language models (PLMs) to automatically generate a question based on each mention. We further proposed a contrastive-learning-based approach to exclude non-quality questions. To assess the quality of sciDataQA, we conducted both automatic and human evaluations, which confirmed the high quality of our dataset.

Based on this dataset, we have developed a novel recursive retrieval approach for scientific dataset recommendation. The key idea of our method is to use UMLS (Bodenreider, 2004), a domain-specific knowledge base, to enrich each question by retrieving relevant background information of a question. Specifically, we construct a terminology tree for each question by expanding each entity into multiple entities that appear in its definition. We then utilized a graph convolutional network to learn the representation of this tree, which integrates information from the original question and relevant background information.

In addition to question answering, we further demonstrated how our dataset could be used to study citation prediction and improve the existing scientific QA systems. In particular, we found that the performance on the low-resource setting of two existing scientific QA systems can be enhanced by fine-tuning them on our dataset, indicating the broad applicability of our dataset. Our contributions can be summarized as follows:

1. *Conceptual:* We propose a novel task and dataset of recommending scientific datasets to answer scientific questions.
2. *Methodological:* We propose a recursive retrieval approach to embed scientific questions.
3. *Application:* We show that our dataset can be used to study citation prediction and improve existing QA systems.

2 sciDataQA dataset

2.1 Collecting scientific datasets

Since there lacks a benchmark that recommends a dataset to a scientific question, we constructed the first scientific dataset recommendation benchmark. In particular, we collected 43,466 datasets from Gene Expression Omnibus (Edgar et al., 2002), where each dataset is a biological data assay. Most of these datasets are gene expression or mutation profiles. Each dataset is a further association with two pieces of text information. One is an author-written summary. The other is the abstract of the corresponding paper that published this dataset. The abstract and summary have 151 and 110 words on average, respectively, which can provide high-quality descriptions for this dataset. Moreover, each dataset is within a large-scale citation network and has on average 24 citations, which can be used as additional context information.

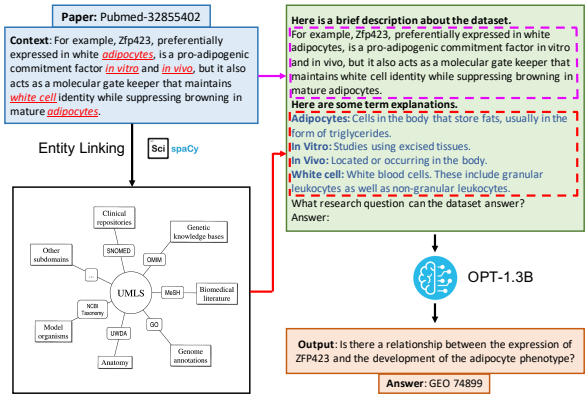

Figure 2: The question generation pipeline for our dataset construction.

2.2 Definition-enriched question generation

Manually creating scientific questions and associating them with scientific datasets require substantial domain experts and cannot be scaled up. As an alternative, we exploited Open Pre-trained Transformer (OPT) (Zhang et al., 2022) to generate questions for each dataset. In particular, we first collected scientific papers that cite a given dataset, assuming that these papers will mention the purposes they use this dataset and these purposes can be converted into high-quality scientific questions. Then for each of these papers, we extracted the sentence that cites the corresponding data. We fed each sentence to OPT as a prompt template to generate the scientific question in **Fig. 2**. Our base prompt is designed by adding "Here is a brief description about the dataset: " before the sentence and adding "What research question can the dataset answer?" after the sentence. To further help PLM better understand scientific text, especially scientific terminology, we developed a definition-enriched prompt. Specifically, we first identified biomedical entities from the sentence and then obtained the definition of these entities from Unified Medical Language System (UMLS) (Bodenreider, 2004). We then appended these definitions to the prompt template. After generating a question from each sentence, we excluded duplicate questions for the same dataset.

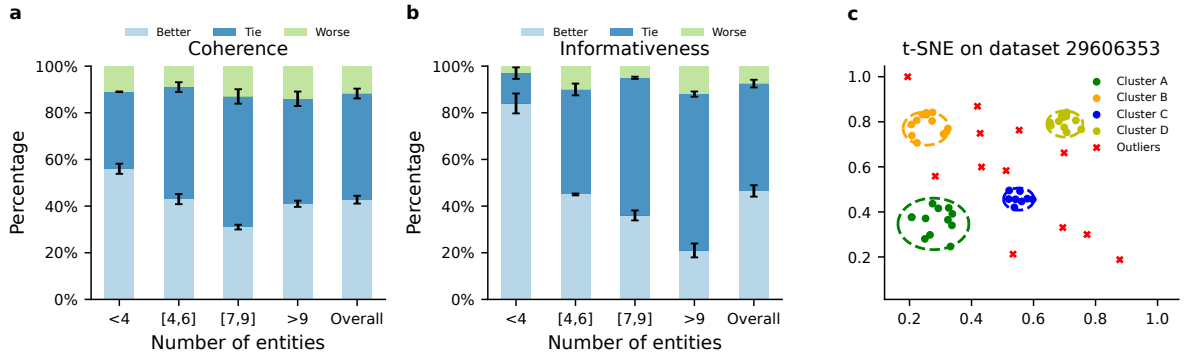

Figure 3: (a) and (b), Comparison in coherence and informativeness between the generation quality with or without enriched definition. Each set contains 125‘ samples. (c), Example of excluding outlier questions.

To validate whether the definition-enriched prompt can improve the quality of questions, we compared the definition-enriched prompt with the base prompt that does not append terminology definitions in **Fig. 3a and Fig. 3b**. In particular, we randomly selected 500 sentences and compared the questions generated by these two prompts. For each pair of questions generated for the sentence, we recruited annotators to assess which question was better in terms of informativeness and coherence. We found that definition-enriched prompts yielded greater or equal coherence on 88.3% of questions and greater or equal to informativeness on 92.5% of questions, indicating the benefits of including definitions. Moreover, we noticed that the improvement of our method is larger when there are fewer entities in the sentences. Since such sentences might be less informative and each entity could play a more important role, augmenting these sentences with definitions could compensate for the sparsity, further confirming the effectiveness of enriching each sentence with definitions.

2.3 Excluding outlier questions using contrastive learning

Intuitively, each dataset should only be able to address a few questions. However, since a dataset might be cited by many papers, we might generate many questions for that dataset. To find the representative question, we used a density-based clustering algorithm OPTICS (Ankerst et al., 1999; Pedregosa et al., 2011) to cluster questions for the same dataset. To obtain feature embeddings for clustering, we applied unsupervised SimCSE (Gao et al., 2021) to the collection of all the questions generated by OPT. Formally, given a BERT-style encoder, we took the last layer hidden-state of `[CLS]` as question representation $\mathbf{h}$ and that with different dropout mask denoted as $\mathbf{h}'$. The training objective for the i-th question can be defined as:

$$\mathcal{L}^i_{\text{unsup}} = -\log \frac{e^{\text{sim}(\mathbf{h}_i,\mathbf{h}'_i)/\tau}}{\sum_{j=1}^{n} e^{\text{sim}(\mathbf{h}_i,\mathbf{h}'_j)/\tau}}, \quad (1)$$

where sim is the cosine similarity, n is the mini-batch size and τ is a temperature hyperparameter. We then excluded questions that are considered outliers by OPTICS. We found that most of these outliers are either not related to or not specified to the corresponding dataset, demonstrating the importance of excluding them from our dataset. To provide a specific explanation, we plot the excluding outcome by t-SNE in **Fig. 3c** on dataset *29606353* and a case study about it in Appendix A.

2.4 Validating question dataset associations

After generating questions for each dataset, we evaluated the question dataset associations. We exploited three evaluation strategies based on existing QA systems, co-citation and manual evaluation.

Evaluating using existing QA systems We first constructed three QA systems based on GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), and UnifiedQA (Khashabi et al., 2020). We then fed a question to each QA system and asked the system to answer this question. Here, each QA system directly provided an answer instead of recommending a dataset. We compared the answer to the summary of each dataset and examined whether the dataset we recommended has higher similarity with this answer generated by existing QA systems, assuming that these systems can partially answer scientific questions. We summarized the results in

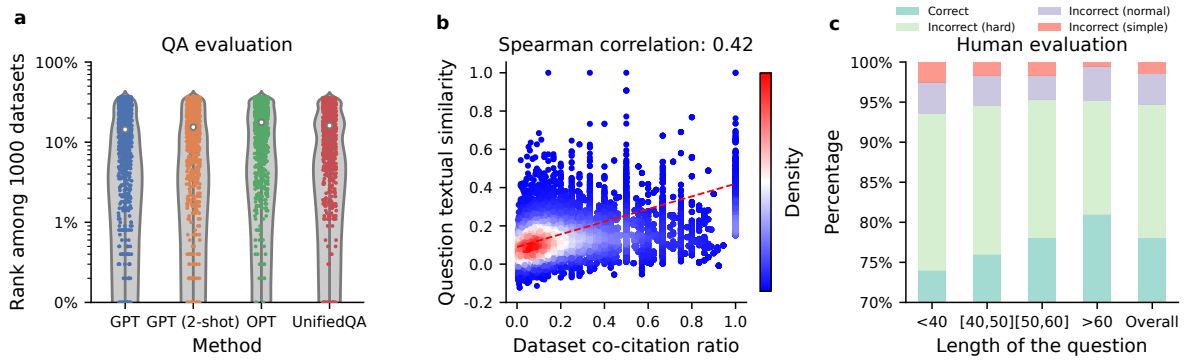

Figure 4: (a), The automatic evaluation on several QA systems. (b), The correlation between the averaged question similarity of two datasets and their co-citation ratio. (c), Human annotation on 1000 samples.

Fig. 4a and observed that 86.4% of the ground-truth datasets are ranked within the top 30.0% among all datasets. This result reflects the substantial consistency between our dataset recommendation and other QA system, supporting the possibility of recommending datasets instead of answering the scientific questions and further suggesting the high-quality associations in sciDataQA.

Evaluating using co-citation relationship We next used the co-citation relationship to evaluate our question dataset associations. We calculate two kinds of similarity metrics between two datasets. The first is a co-citation similarity based on how many papers cite them using Jaccard similarity. The second is the semantic similarity between the generated questions of the two datasets. We found that these two similarity metrics are highly consistent with a Spearman correlation of 0.42 (**Fig. 4b**). As co-citation similarity has been extensively used to measure scientific paper similarity(Boyack et al., 2013), this high consistency indicates that we generated similarity questions for similar datasets, further confirming the quality of the associations.

Manual evaluation The above two large-scale automatic evaluations demonstrate the quality of our question data associations. We next conducted a manual evaluation by designing multiple-choice questions. In particular, for each dataset, we provided four questions: the ground truth associated question (positives), a simple negative question whose representation is the farthest from the positive question (simple negatives), a normal negative question randomly sampled from another dataset (normal negatives), and a hard negative question that is generated from another sentence that in the same paper paragraph as the positive question (hard negatives). We asked the human annotators to select the question that best matched the dataset according to the dataset summary.

We found that human annotators achieved 78% accuracy in this multiple-choice-based evaluation (**Fig. 4c**), indicating that our dataset question associations are consistent with human knowledge. We noted that most of the incorrectness fell into the hard negatives category. These hard negatives require the most domain knowledge compared to simple negative and normal negatives since sentences of these hard negatives are in the same paragraph with positive questions.

3 Recursive definition retrieval for dataset recommendation

3.1 Problem definition

Given a scientific question $Q = (q_1, q_2, \ldots, q_i)$, and a set of datasets $\mathcal{D} = (D_1, D_2, \ldots, D_T)$, we aim to select the best-matched dataset for that question. For each dataset D_t, we also have a dataset summary $S_t = (s_1^t, s_2^t, \ldots, s_m^t)$ and a scientific paper abstract $A_t = (t_1^t, t_2^t, \ldots, t_n^t)$ describing that dataset. We do not consider citation networks for the recommendation.

3.2 Recursive definition retrieval

Scientific questions might contain terminologies that cannot be easily processed by PLMs (Lavrenko and Croft, 2017; Yu et al., 2021). Motivated by the promising results of enriching questions with definitions in dataset generation, we also propose to include definitions for each question for a better recommendation. Here, we propose a recursive definition retrieval approach that recursively expands

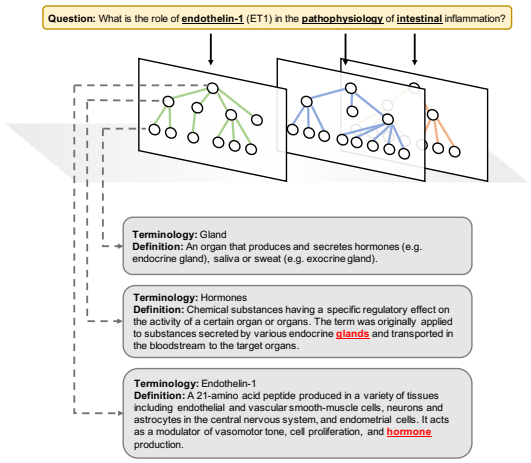

Figure 5: An question example processed by the recursive retrieval approach. There are three identified entities in the question and we constructed an entity tree for each entity based on their definition texts.

a question into an entity tree.

Superficially, for each question Q, we convert it to a multi-root entity tree. The roots of this tree are entities in this question identified through entity linking. We then obtain the definition of each entity and find new entities in the definition using ScispaCy (Neumann et al., 2019). These new entities will be inserted into this tree as child nodes. We recursively repeat this process by expanding more layers in this tree, where a child entity is mentioned in the definition of the parent entity. To prevent very deep and large trees that could be computationally intensive, we will terminate the expansion if the new entity is very different from its corresponding root entity based on the definition textual similarity. An example of the multi-root entity tree is shown in **Fig. 5**. We set the maximum depth of the tree and similarity threshold as hyperparameters. This process will help us enrich the question Q by augmenting it with related definitions.

3.3 Tree-augmented question embedding

Each node in the tree is associated with an entity and a definition. We can now use them to augment the original question. To achieve this, we learn two graph convolutional networks (GCN) (Defferrard et al., 2016) to embed entity names and entity definitions respectively: Initial node features for GCN are the BERT embedding of entity names or definitions. We then separately aggregated the entity-based embedding of all roots and the definition-based embeddings of all roots. Instead of aggregating the embeddings of all nodes, we only consider roots, which are entities in the original question. This design enables conservatively enriching the question without adding too much irrelevant information. These two aggregated embeddings are concatenated with question embeddings to get the final representation.

4 Experimental results

We randomly select 10000 samples from sciDataQA and split them as training (80%), dev (10%), and test set (10%), using cross-validation. The hyper-parameter selection is presented in Appendix B. We compare our method with two text classification models: CLEncoder (Gao et al., 2021) and UniEncoder (Devlin et al., 2018). Since neither of them retrieve and augment extra information for the question, our comparison can show the importance of tree-based question augmentation.

CLEncoder. Wu et al. (2022) point out that when one passage could be the positive passage of multiple questions, there would be a higher probability of the passage appearing as both positive and negative instances in one batch simultaneously. Therefore, we design the negative instance, in addition to the in-batch negative technique, to address this issue. Specifically, for each question x_i, we take the abstract and summary of its corresponding dataset as positive instance x_i^+ and that of a randomly chosen dataset other than the positive dataset as negative instance x_i^-, then conduct in-batch contrastive learning. The training objective $\mathcal{L}_{\text{sup}}^i$ is formulated as:

$$-\log \frac{e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N}(e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_j^+)/\tau}+e^{\text{sim}(\mathbf{h}_i,\mathbf{h}_j^-)/\tau})}, \quad (2)$$

where symbols are defined the same as Eq. 1.

UniEncoder. We also use a BERT encoder with a binary classification layer to do our task. Formally, we set the concatenation of the question, dataset summary, and dataset abstract with different separate tokens as inputs of the BERT encoder and use the [CLS] token's hidden state as the input of the classification layer.

Main Results. According to Table 1, we can see our approach substantially outperforms the baselines on a variety of top K accuracy, indicating the effectiveness of augmenting the question with entity definitions. We further noticed that the improvement is larger when K is smaller. For example, our method achieves 7.3% enhancement

	ACC@1	ACC@5	ACC@10	ACC@20	ACC@50
CL	0.146	0.302	0.396	0.485	0.618
CL + GCN	0.151	0.316	0.400	0.487	0.618
Uni	0.102	0.293	0.383	0.494	0.650
Uni + GCN	**0.175**	**0.343**	**0.434**	**0.535**	**0.661**

Table 1: The accuracy of our model on the dataset, where CL denotes CLEncoder, and Uni denotes UniEncoder. Test on KRISSBERT.

	ACC@1	ACC@5	ACC@10	ACC@20	ACC@50
BERT	0.092	0.239	0.314	0.386	0.533
SciBERT	0.151	0.316	0.398	0.495	0.623
PubMedBERT	0.168	0.332	**0.445**	0.531	0.653
KRISSBERT	**0.175**	**0.343**	0.434	**0.535**	**0.661**

Table 2: Performances of different pre-trained models. Test on UniEncoder + GCN.

on top 1 accuracy, which is much higher than the 4.1% enhancement on top 20 accuracy, when compared to UniEncoder. However, for CLEncoder, our approach only has little improvement over the baseline. We assume that in UniEncoder, the terminology tree could interact with both the question and abstract/summary. Through the self-attention network, the connection between the specific entity in the question and our dataset could be further considered for recommendation.

Pre-trained Models. In this experiment, we use BERT (Devlin et al., 2018), SciBERT (Beltagy et al., 2019), KRISSBERT (Zhang et al., 2021), and PubMedBERT (Gu et al., 2020)'s pre-trained weights to initialize our encoders. And we investigate the effect of different pre-training parameters on model performance.

According to Table 2, the results show that the SciBERT, KRISSBERT, and PubMedBERT all outperform the original BERT by a large margin. An apparent reason is that the original BERT is not good at processing biomedical data. And among the three BERT, we can see the KRISSBERT outperforms the other two BERT. One possible explanation is that the KRISSBERT uses PubMedBERT's parameters and is continuously fine-tuned on the UMLS dataset, which is also the data source for our tree construction.

Ablation Studies. We construct a terminology tree to enrich the information of specific questions by recursive retrieval. To analyze how the variables in the tree influence the results, we conduct detailed ablation studies (**Fig. 6a**).

We first modify the tree depth from 2 to 5 when the similarity threshold is fixed at 0.8. The overall trend shows that when the similarity threshold is constant, as the tree depth increases, the recommendation's accuracy is better. We assume that the tree could provide more details to understand the given questions when it gets deeper. However, when the depth reaches 5, more irrelevant information will harm the model performance.

Then, we set the depth to 4 and tune the similarity threshold among 0.8, 0.9, 0.95, and 0.99. Our method becomes less accurate as the threshold keeps increasing. This indicates that too large a threshold will decrease the scale of the tree structure and thus can't provide enough information for an accurate recommendation.

5 Applications of sciDataQA

In addition to the main application of dataset recommendation, sciDataQA can also be used for other applications involving scientific datasets. We investigated two such applications here and raised more applications in the Future Work section.

5.1 Providing additional training data for existing QA systems

First, sciDataQA can be used to fine-tune existing QA systems for scientific question answering. In particular, we can treat the question and the summary of its corresponding dataset as a question-answer pair. We can obtain 7500 such pairs from our training set. We can then exploit these questions to fine-tune existing QA systems (Yoo et al., 2021; Wang et al., 2021b; Meng et al., 2022; Ye et al., 2022). To validate this application, we studied two QA approaches based on UnifiedQA (BART) (Khashabi et al., 2020) and Instruction Tuning (T5) (Sanh et al., 2022). We evaluate their performance in the few-shot setting on an independent scientific QA dataset ScienceQA (Lu et al., 2022) (biology subset). Both approaches are fine-tuned using the entire sciDataQA.

We found that our dataset substantially improved the performance of both QA systems (**Fig. 6c**). For example, UnifiedQA fine-tuned on our dataset obtained a 62.38 accuracy in the zero-shot setting, which is much higher than the 58.63 accuracy using UnifiedQA only. The improvement is more significant when there are less training data from ScienceQA, especially in the zero-shot scenario,

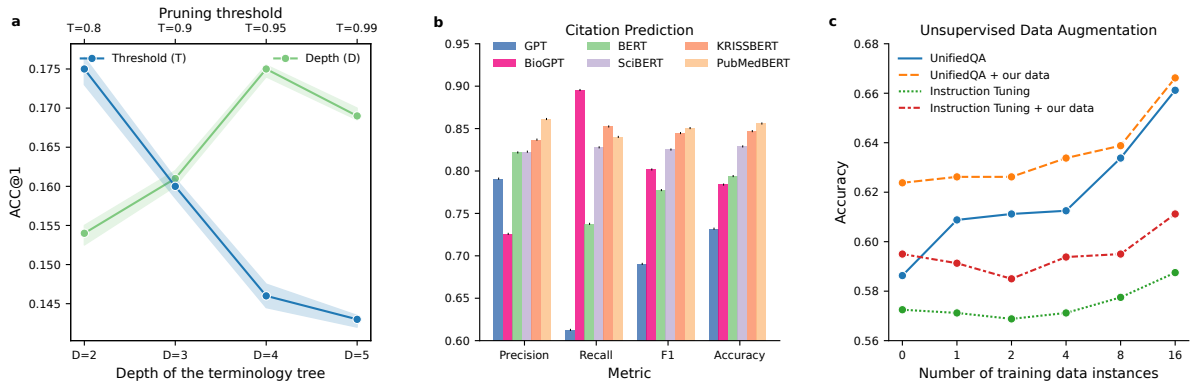

Figure 6: (a), Ablation studies on the dataset recommendation task. Test on KRISSBERT. (b), The citation prediction. (c), Comparison between with or without our data as unsupervised data augmentation on ScienceQA.

further indicating the advantage of leveraging sciDataQA as additional training data.

5.2 Citation Prediction

Moreover, our dataset can be used to study citation prediction. Citation prediction is an important task in scientific literature analysis (Bai et al., 2019). It aims to predict the future citation relationship between papers, which has critical implications for detecting emerging research problems and improving scientific paper writing efficiency.

There exists a substantial amount of citation relationships in our dataset, which can be used to predict and evaluate the citation prediction. Different from existing citation prediction datasets (Cohan et al., 2020), sciDataQA focuses on recommending citation of dataset papers. As a result, we can additionally consider the dataset summary as a feature. Specifically, given two dataset papers and their summaries, we will predict whether one paper cites the other. As two papers that have similar summaries are more likely to cite each other, we concatenated their summaries as input and trained a binary classifier. We considered encoders based on BERT (Devlin et al., 2018), PubMedBERT (Gu et al., 2020), KRISSBERT (Zhang et al., 2021), and, SciBERT (Beltagy et al., 2019). We considered decoders based on GPT-2 (Radford et al., 2019) and BioGPT (Luo et al., 2022) to predict the citation.

We summarized the results in **Fig. 6b**. We observed that all these PLMs achieved in general high prediction results, supporting the high quality of our dataset. Moreover, we observed a noticeable discrepancy among these PLMs. In particular, domain-specific language models, such as PubMedBERT, KRISSBERT, and BioGPT, perform better than general language models, such as GPT and BERT. This observation is consistent with previous works (Gu et al., 2020) that domain-specific language models have better performance on a variety of downstream applications. Thus, our dataset also offers an application to compare various of PLMs.

6 Related Work

6.1 Dataset generation using language models

Existing approaches to dataset generation mainly focus on fine-tuning the generative models using existing training data and then generating additional training data (Anaby-Tavor et al., 2020; Kumar et al., 2020; Puri et al., 2020; Lee et al., 2021; He et al., 2021; Vu et al., 2021; Mekala et al., 2022). The generative dataset augmentation has been applied to a variety of applications, including question answering (Alberti et al., 2019), commonsense reasoning (Yang et al., 2020), semantic textual similarity (Schick and Schütze, 2021), labeled documents (Mekala et al., 2021), biomedical factoid question answering (Pappas et al., 2022), and query reformulations (Adolphs et al., 2022). Recently, SuperGen (Meng et al., 2022) and ZeroGen (Ye et al., 2022) generate training data guided by label-descriptive prompts. Here, we generate questions for the scientific dataset. There are two major differences between our work and existing approaches. First, we focus on a novel application of generating scientific questions for scientific dataset recommendation. Second, instead of fine-tuning the large language model using training data, we utilize background definition information to prompt the language model without using any training data.

6.2 Scientific question answering

Scientific question answering is a challenging task that has been studied in both single text modality (Khashabi et al., 2018; Clark et al., 2018; Mihaylov et al., 2018; Khot et al., 2020; Lu et al., 2022) and multi-modal reasoning (Krishnamurthy et al., 2016; Kembhavi et al., 2016, 2017; Kafle et al., 2018; Sampat et al., 2020; Lu et al., 2021a,b). To leverage the reasoning path for constructing better QA systems, enhanced datasets (Jansen et al., 2018; Jhamtani and Clark, 2020; Dalvi et al., 2021) annotate explanations for the question-answer pairs from the perspective of explanation graphs, reasoning chains, and entailment trees respectively. To construct scientific question-answering systems, previous approaches have exploited K-nearest neighbour (Altman, 1992), latent dirichlet allocation (Blei et al., 2003), the co-authors' network (Luong et al., 2012), writing style (Yang and Davison, 2012), citations (Küçüktunç et al., 2012), and PLMs (Khashabi et al., 2020; Xu et al., 2021; Huang et al., 2022) to perform the answer recommendation or generation on scientific papers. By contrast, we don't provide the answer explanations explicitly, but recommend a dataset for scientists to study in order to answer this question.

6.3 Dataset recommendation

There are two scenarios of dataset recommendation: 1) recommendation based on user query (Leme et al., 2013; Ben Ellefi et al., 2016; Patra et al., 2020; Singhal et al., 2013; Altaf et al., 2019); 2) recommendation based on provided dataset (Wang et al., 2021a). These recommendation studies focused on computer science instead of the scientific field and have never been applied to the rich collection of Gene Expression Omnibus. To fill in this gap, we provide a high-quality dataset and novel methods for scientific dataset recommendation.

7 Discussions and Future Work

Dataset generation with PLM. One of our key contributions is to use the pre-trained language model to generate specific questions. We found that the design of prompts for language models is essential for the quality of our questions. Specifically, if we change the order of background information and the dataset description, the quality of the generated questions will be lower, and the model might not generate anything for some datasets. As a result, designing a reasonable and effective prompt is critical for a PLM to generate high-quality questions. Moreover, the enriched definitions have been demonstrated to be essential for question generation. However, we also observed that adding too much background might hurt the generation's performance by introducing irrelevant information. In the future, we want to develop a better approach to incorporate background information into pre-trained language models for knowledge-aware generation.

Recursive retrieval for dataset recommendation We have proposed an entity-tree-based approach for dataset recommendation. Currently, we need to limit the number of nodes in the tree by using a pruning algorithm. Without this constraint, the number of nodes in the tree grows exponentially with increasing depth, and the memory usage will influence the training and inference process seriously. In this work, we set the similarity threshold statically, which proves effective in the recommendation. However, to get a better understanding of each question, it may need information with different granularities for different kinds of entities. We leave the exploration of dynamic pruning algorithms as future work for better scale control in the entity tree.

8 Conclusion

In this paper, we study a novel problem of scientific dataset recommendation via our proposed dataset, sciDataQA. We argue that instead of answering challenging scientific questions directly, it is more realistic to recommend a scientific dataset that might be able to solve this question. To construct our dataset, we developed a novel definition-enriched approach to generate high-quality scientific questions using a pre-trained language model OPT. Both automatic and human evaluations confirm the quality of our dataset.

Based on sciDataQA, we developed a tree-augmented recursive retrieval dataset recommendation method and obtained substantial improvement on several strong baselines. We further demonstrated how our dataset could be exploited to recommend scientific citations and improve existing scientific QA systems. Collectively, we have proposed a comprehensive solution for scientific dataset recommendation, including defining the task, building a new dataset, and proposing the novel recommendation method.

Limitations

Due to the limitation of computing resources, one deficiency of this work is that our dataset was generated with OPT-1.3B, whose parameter is much smaller than the popular GPT-3 or a pre-trained model of equivalent capability. However, with proper data filtering algorithms, the promising results of our recommendation method and downstream applications showed that our dataset is of high quality, which confirms the validity of our approach.

Another limitation is that we don't manually annotate the gold answer for each question in our dataset because of the high cost of professional human resources. Since our primary goal is to build a dataset recommendation system to solve challenging science QA, standard answers seem less necessary. Furthermore, treating the dataset summary's first sentence as the answer is proven to be an effective workaround, as pre-training on it significantly improves the QA system's accuracy on ScienceQA.

Acknowledgements

References

Leonard Adolphs, Michelle Chen Huebscher, Christian Buck, Sertan Girgin, Olivier Bachem, Massimiliano Ciaramita, and Thomas Hofmann. 2022. Decoding a neural retriever's latent space for query suggestion.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Basmah Altaf, Uchenna Akujuobi, Lu Yu, and Xiangliang Zhang. 2019. Dataset recommendation via variational graph autoencoder. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 11–20. IEEE.

Naomi S Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7383–7390.

Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. 1999. Optics: Ordering points to identify the clustering structure. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, page 49–60, New York, NY, USA. Association for Computing Machinery.

Xiaomei Bai, Fuli Zhang, and Ivan Lee. 2019. Predicting the citations of scholarly paper. *Journal of Informetrics*, 13(1):407–418.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Mohamed Ben Ellefi, Zohra Bellahsene, Stefan Dietze, and Konstantin Todorov. 2016. Dataset recommendation for data linking: An intensional approach. In *European Semantic Web Conference*, pages 36–51. Springer.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Kevin W Boyack, Henry Small, and Richard Klavans. 2013. Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, 64(9):1759–1767.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL*.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29.

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

Ron Edgar, Michael Domrachev, and Alex Lash. 2002. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. nucl acids res 30: 207-210. *Nucleic acids research*, 30:207–10.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2021. Generate, annotate, and learn: Nlp with synthetic text. *CoRR*, abs/2106.06168.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. 2022. Clues before answers: Generation-enhanced multiple-choice QA. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3272–3287. Association for Computational Linguistics.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A corpus of explanation graphs for elementary science questions supporting multi-hop inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kushal Kafle, Scott Cohen, Brian Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *CVPR*.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

D. Khashabi, S. Min, T. Khot, A. Sabhwaral, O. Tafjord, P. Clark, and H. Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090.

Jayant Krishnamurthy, Oyvind Tafjord, and Aniruddha Kembhavi. 2016. Semantic parsing to probabilistic programs for situated question answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 160–170, Austin, Texas. Association for Computational Linguistics.

Onur Küçüktunç, Erik Saule, Kamer Kaya, and Ümit V Çatalyürek. 2012. Recommendation on academic networks using direction aware citation analysis. *arXiv preprint arXiv:1205.1143*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Victor Lavrenko and W. Croft. 2017. Relevance-based language models. *ACM SIGIR Forum*, 51:260–267.

Kenton Lee, Kelvin Guu, Luheng He, Timothy Dozat, and Hyung Won Chung. 2021. Neural data augmentation via example extrapolation. *ArXiv*, abs/2102.01335.

Luiz André P Paes Leme, Giseli Rabello Lopes, Bernardo Pereira Nunes, Marco Antonio Casanova, and Stefan Dietze. 2013. Identifying candidate datasets for data interlinking. In *International Conference on Web Engineering*, pages 354–366. Springer.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021a. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *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-IJCNLP 2021)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. 2012. Publication venue recommendation using author network's publication history. In *Asian Conference on Intelligent Information and Database Systems*, pages 426–435. Springer.

Dheeraj Mekala, Varun Gangal, and Jingbo Shang. 2021. Coarse2Fine: Fine-grained text classification on coarsely-grained annotated data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 583–594, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dheeraj Mekala, Tu Vu, Timo Schick, and Jingbo Shang. 2022. Leveraging qa datasets to improve generative data augmentation.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutsopoulos. 2022. Data augmentation for biomedical factoid question answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 63–81, Dublin, Ireland. Association for Computational Linguistics.

Braja Gopal Patra, Kirk Roberts, and Hulin Wu. 2020. A content-based dataset recommendation system for researchers—a case study on gene expression omnibus (geo) repository. *Database*, 2020.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shailaja Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-linguistic question answering (vlqa) challenge.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ayush Singhal, Ravindra Kasturi, Vidyashankar Sivakumar, and Jaideep Srivastava. 2013. Leveraging web intelligence for finding interesting research datasets. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 321–328. IEEE.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

X Wang, F van Harmelen, and Z Huang. 2021a. Biomedical dataset recommendation. In *10th International Conference on Data Science, Technology and Applications, DATA 2021*, pages 192–199. SciTePress.

Zhen Wang. 2022. Modern question answering datasets and benchmarks: A survey.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021b. Towards zero-label language learning.

Bohong Wu, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2022. Sentence-aware contrastive learning for open-domain passage retrieval. In *The 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*.

Weiwen Xu, Yang Deng, Huihui Zhang, Deng Cai, and Wai Lam. 2021. Exploiting reasoning chains for multi-hop science question answering. pages 1143–1156.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Zaihan Yang and Brian D Davison. 2012. Venue recommendation: Submitting your paper with style. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 681–686. IEEE.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. Zerogen: Efficient zero-shot learning via dataset generation.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving query representations for dense retrieval with pseudo relevance feedback.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Knowledge-rich self-supervision for biomedical entity linking.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

A Dataset Examples

According to Table 3, the questions in the clusters represent different aspects of the dataset (in this case, BCL11A's regulation, expression, structure, and function), and their answers can be found in the abstract and summary. However, the outlier points are either too generic (Outlier 1) or misled by the entity (Outlier 2) that is peripheral in the context.

B Implementation Details

For dataset recommendation, we fix batch size as 16, learning rate as 1e-5, and train epochs as 10. For citation prediction, we fix batch size as 32, learning rate as 2e-5, and train epochs as 10. For SimCSE (Gao et al., 2021), UnifiedQA (Khashabi et al., 2020), and Instruction Tuning (Sanh et al., 2022), we use their original training scripts with no modification in hyperparameters.

Abstract	Fetal hemoglobin (HbF, $\alpha 2\gamma 2$) level is genetically controlled and modifies severity of adult hemoglobin (HbA, $\alpha 2\beta 2$) disorders, sickle cell disease and β-thalassemia. Common genetic variation affects expression of BCL11A, a regulator of HbF silencing. To uncover how BCL11A supports the developmental switch from γ- to β- globin, we use a functional assay and protein binding microarray to establish a requirement for a zinc-finger cluster in BCL11A in repression, and identify a preferred DNA recognition sequence. This motif appears in embryonic and fetal-expressed globin promoters, and is duplicated in γ-globin promoters. The more distal of the duplicated motifs is mutated in individuals with hereditary persistence of fetal hemoglobin. Using the CUT&RUN approach to map protein binding sites in erythroid cells, we demonstrate BCL11A occupancy preferentially at the distal motif, which can be disrupted by editing the promoter. Our findings reveal that direct γ-globin gene promoter repression by BCL11A underlies hemoglobin switching.
Summary	Fetal hemoglobin (HbF) level is genetically controlled and modifies severity of adult hemoglobin (HbA) disorders. Common genetic variation affects expression of BCL11A, a critical regulator of HbF silencing. Current models suggest that BCL11A acts at a distance from the gamma-globin genes via long-distance chromosomal interactions. Here we use a functional cellular assay and protein-binding microarray to establish a requirement for a zinc-finger cluster of BCL11A for globin repression, and identify a preferred DNA recognition sequence (TGACCA). The motif is present in embryonic and fetal-expressed globin promoters, and duplicated in gamma-globin promoters, yet only the distal motif is mutated in alleles of individuals with hereditary persistence of hemoglobin. Using CUT&RUN to map protein binding sites, we detected BCL11A occupancy preferentially at the distal motif, and validated its absence in HbF-expressing, promoter-edited erythroid cells. Taken together, our findings reveal that direct gamma-globin gene promoter repression by BCL11A underlies hemoglobin switching.
Cluster A	How is the Bcl11a gene regulated?
Cluster B	What is the relationship between the expression levels of Bcl11a and the transcriptional activity of the human hematopoietic stem cell (HSC) lineage?
Cluster C	What is the structure of the zinc finger domain of BCL11A?
Cluster D	How does the gene expression profile change in response to the presence or absence of Bcl11a?
Outlier 1	What is the average length of a DNA fragment?
Outlier 2	What is the role of β-Globin in the regulation of gene expression?

Table 3: Case study of excluding outlier questions on dataset *29606353*.