CQSumDP: A ChatGPT-Annotated Resource for Query-Focused Abstractive Summarization Based on Debatepedia

Md Tahmid Rahman Laskar[1,3], Mizanur Rahman[2,3], Israt Jahan[3], Enamul Hoque[3], Jimmy Huang[3,*]
[1]Dialpad Canada Inc., [2]Royal Bank of Canada, [3]York University[†]
Toronto, Ontario, Canada
{tahmid20,mizanurr,israt18,enamulh,jhuang}@yorku.ca

Abstract

Debatepedia is a publicly available dataset consisting of arguments and counter-arguments on controversial topics that has been widely used for the single-document query-focused abstractive summarization task in recent years. However, it has been recently found that this dataset is limited by noise and even most queries in this dataset do not have any relevance to the respective document. In this paper, we present a methodology for cleaning the Debatepedia dataset by leveraging the generative power of large language models to make it suitable for query-focused abstractive summarization. More specifically, we harness the language generation capabilities of ChatGPT to regenerate its queries. We evaluate the effectiveness of the proposed ChatGPT annotated version of the Debatepedia dataset using several benchmark summarization models and demonstrate that the newly annotated version of Debatepedia outperforms the original dataset in terms of both query relevance as well as summary generation quality. We will make this annotated and cleaned version of the dataset publicly available.

1 Introduction

Abstractive summarization is a natural language processing technique that involves generating a concise and coherent summary of a longer piece of text while preserving its most important information (Yao et al., 2017). Query-focused abstractive summarization is a specific type of abstractive summarization that generates a summary of the given text that is tailored to a specific query or topic of interest (Baumel et al., 2018; Goodwin et al., 2020; Su et al., 2020; Xu and Lapata, 2021; Laskar et al., 2020a,b, 2022d). In other words, the summary is focused on answering a specific question or addressing a particular topic, rather than providing a general overview of the text. One widely used dataset for this task is the Debatepedia[1] dataset that consists of arguments and counter-arguments on conversational topics (Nema et al., 2017).

The query-focused summarization of argumentative text is a challenging task that has gained increasing attention in recent years due to its potential applications in various domains, such as policy-making, journalism, and legal reasoning (Nema et al., 2017; Laskar et al., 2020a). However, it has been recently found that the quality of the Debatepedia dataset that is widely used for the query-focused abstractive summarization task is limited by noise, with many of the queries in this dataset does not have any relevance with the source document (Laskar et al., 2022d). Since Debatepedia is a rich source of argumentative text on controversial topics that can serve as a valuable resource for developing and evaluating summarization models, in this paper, we present a novel methodology to annotate the Debatepedia dataset to make it a useful resource for query-focused abstractive summarization. Our data annotation approach leverages the language modeling (Radford et al., 2019) capabilities of ChatGPT[2], a large pre-trained language model (Devlin et al., 2018; Brown et al., 2020; Ouyang et al., 2022) that has shown an impressive capability of generating fluent and coherent text (Qin et al., 2023; Bang et al., 2023; Yang et al., 2023; Kuzman et al., 2023; Gao et al., 2023; Wang et al., 2023; Zhou et al., 2023; Kocoń et al., 2023; Kocmi and Federmann, 2023). Using ChatGPT as the annotator, we regenerate the queries in the Debatepedia dataset to remove the noise in this dataset. We validate the effectiveness of our methodology by conducting extensive experiments on our newly constructed dataset that leverages ChatGPT as the annotator. Our major contributions in this paper

*Contact Author.
[†]All work being done at York University.

[1]https://github.com/PrekshaNema25/DiverstiyBasedAttentionMechanism
[2]https://openai.com/blog/chatgpt/

are summarized below:

- We proposed a novel methodology for cleaning and annotation of the Debatepedia dataset using a large language model, i.e., ChatGPT to improve its suitability for query-focused abstractive summarization. This paper also opens up a promising avenue to utilize ChatGPT as the annotator for other tasks beyond text summarization that can significantly reduce the overall cost of data annotation.
- We conducted extensive experiments using benchmark summarization models on our ChatGPT-annotated cleaned version of Debatepedia for Query-Focused Abstractive Summarization and observe that it outperforms the original dataset in terms of both query relevance and summary generation quality.
- Our annotated dataset will be made publicly available such that it can serve as a valuable resource for further research on query-focused abstractive summarization.

2 Related Work

Query-focused abstractive summarization using neural models has gained increasing attention in recent years (Baumel et al., 2018; Laskar et al., 2022d). The recent success of transformer-based encoder-decoder models (Liu and Lapata, 2019; Lewis et al., 2019; Raffel et al., 2019; Zhang et al., 2019) on generic[3] abstractive summarization has also inspired researchers to utilize such models for query-based abstractive summarization (Goodwin et al., 2020; Vig et al., 2021; Laskar et al., 2020a,b, 2022d), leading to state-of-the-art performance in benchmark query-based summarization and answer generation datasets, such as DUC[4] (Feigenblat et al., 2017; Roitman et al., 2020; Xu and Lapata, 2021, 2020), AQuaMuSe (Kulkarni et al., 2020), QMSum (Zhong et al., 2021), WikiHowQA (Deng et al., 2019), PubMedQA (Jin et al., 2019), MediQA (Savery et al., 2020), MS-MARCO (Wang et al., 2018), Debatepedia (Nema et al., 2017), etc. Though some studies (Abdullah and Chali, 2020) also attempted to generate the queries in generic summarization datasets (e.g., CNNDM (Nallapati et al., 2016)) using the source document and the reference summary to enable such datasets for query-focused summarization, we find that these queries are generated by directly extracting words or tokens from the reference summaries. As a result, the summarization models have unexpected access to the keywords in the gold reference summaries.

Among the datasets mentioned above, DUC and AQuaMuSe require generating summaries from multiple documents, usually from the news domain. The QMSum dataset is proposed for query-based meeting summarization, while WikiHowQA is constructed from the WikiHow knowledgebase and used for answer summary generation for questions that start with "How to". Meanwhile, PubMedQA and MediQA datasets are constructed from the biomedical domain. One notable exception among these datasets is the Debatepedia dataset since it requires generating abstractive summaries from a short document containing argumentative text. None of the other datasets mentioned above addressed the issue of generating query-based summaries from documents containing arguments and counter-arguments. This makes Debatepedia a great resource for researchers to develop methods to summarize a short document containing argumentative text for the given query.

However, it has been found recently that many samples in the Debatepedia dataset are not actually query oriented (Laskar et al., 2022d). Moreover, it was also observed that fine-tuning pre-trained neural models in this dataset without considering the query incorporation could achieve almost similar performance as the query-focused summarization models (Laskar et al., 2022d). Thus, there remains a scarcity of datasets specifically tailored for creating condensed summaries of argumentative texts that are relevant to a single query.

To address the above issue, in this work, we seek to clean the Debatepedia dataset to make it usable for query-focused single document abstractive summarization of argumentative text. For that purpose, we propose a novel methodology that leverages the text generation capability of prompt-based language models (Liu et al., 2023; Ouyang et al., 2022; Brown et al., 2020). To this end, we utilize ChatGPT, a powerful generative Large Language Model (LLM) developed by OpenAI[5] which has received a lot of attention recently due to its impressive

[3] In Generic Abstractive Summarization, the summaries are generated based on only the given source document.

[4] https://duc.nist.gov/data.html

[5] https://openai.com/

Example 1: *Query having no relevance with the document and the summary.*
Query: Does an MBA enhance leadership skills?
Document: Business schools might improve your quantitative presentation and communication skills. It might but get you thinking about ethical and strategy. But two years of case studies aren't go to turn you into a leader if you weren't died one. There's no learning charisma persuasiveness elegance or gut instinct.
Reference Summary: PhD will not improve cm factors of leaders.
Example 2: *One word summary having no relevance with the query or document.*
Query: Education : do child benefit from watching tv?
Document: by watching news child can learn about geography politics advances in science – everything simply and later explained . furthermore child learn about real-life situation that happens on everyday basis which will benefit them in the future.
Reference Summary: News.
Example 3: *The length of the summary is longer than the document with the query being irrelevant.*
Query: activists : where do the keys activists and organizations stand ?
Document: see an analyses of the article ...
Reference Summary: philip martin of berkeley davis and michael teitelbaum the mirage of mexican guest workers nov/dec # foreign affairs .
Example 4: *More of a close-ended question.*
Query: friendships : does twitter harms relationships ?
Document: twitter helps those stay in touches no matter how far they may be from each other .
Reference Summary: long-distance friendships .

Table 1: Some examples demonstrating the limitations in the Debatepedia dataset.

language generation capability – ensuring high fluency, coherence, and grammatical correctness on its generated texts (Qin et al., 2023). ChatGPT like such Generative LLMs (Scao et al., 2022; Tay et al., 2022; Thoppilan et al., 2022; Fedus et al., 2021; Hoffmann et al., 2022; Zeng et al., 2022; Chowdhery et al., 2022; Sanh et al., 2021a) that leverage the prompt-based learning mechanism have obtained impressive performance in few-shot and zero-shot learning scenarios, inspiring researchers to also explore some new applications of these models, such as data annotation (Wang et al., 2021; Ding et al., 2022). In this paper, we also harness the text generation power of ChatGPT to fix the queries in the Debatepedia dataset to construct a cleaned version of the dataset that could be used for query-focused abstractive summarization of argumentative text. With extensive experiments, we validate that our proposed cleaned version of the Debatepedia dataset overcomes the limitations of the existing noisy version of this dataset.

3 Debatepedia Dataset Limitations

Debatepedia is a publicly available dataset of arguments and counter-arguments on debate topics, proposed by Nema et al. (Nema et al., 2017). It contains 13,573 query-document-summary pairs. The average number of words per document, summary, and query in the Debatepedia dataset is 66.4, 11.16, and 9.97, respectively. The dataset covers a wide range of topics, such as politics, sports, and technology, and has been extensively used in recent years to build query-based summarization models for argumentative text.

However, the quality of Debatepedia as a dataset for query-based summarization has lots of limitations (see Table 1 for some examples), as it has been found recently that many queries in this dataset are not relevant to the document (Laskar et al., 2022d). Based on a randomly sampled 100 instances, it has been found in a recent study (Laskar et al., 2022d) that:

- 52% of the queries in this dataset have no relevance to the documents or the summaries, as demonstrated in Table 1.
- 70% of the queries are close-ended (i.e., Yes/No type) questions (see Example 4 in Table 1).
- Though, many queries in this dataset are relevant to the documents but the summaries

are more of generic due to shorter document length. Note that the average size of the document in this dataset is only 66.4 words on average.

In addition, many instances in this dataset only contain one word summary (see Example 2 in Table 1) for a given query that appears both in the training and evaluation sets, which may also help the model to memorize such words for similar queries during the training phase. These issues may lead to an unexpected increase in the ROUGE score when the model starts learning to reproduce those words in the summary during the evaluation phase. Furthermore, we also find some instances where the length of the summary is longer than the document length, which usually happens in short documents (see Example 3 in Table 1).

To address these limitations, we propose a methodology for cleaning the Debatepedia dataset via leveraging ChatGPT as the data annotator to regenerate the queries. In the following, we describe our methodology.

4 Our Annotation Methodology

The recently released ChatGPT model has demonstrated impressive performance to solve a wide-range of problems, from generating fluent and coherent summaries from documents to solving mathematical problems, along with solving challenging information retrieval tasks, such as open-domain question answering, neural machine translation, writing programming solutions, and etc (Qin et al., 2023; Guo et al., 2023). In this work, we leverage ChatGPT as the annotator to fix the issues in the Debatepedia dataset to use it for query-focused abstractive summarization. We denote our **C**hatGPT annotated cleaned dataset for **Q**uery Focused Abstractive **Sum**marization based on **D**ebate**p**edia as the **CQSumDP** dataset.

As demonstrated in the previous section the Debatepedia dataset has several limitations, containing noisy and irrelevant contents (e.g., queries/documents/summaries). To address these issues, we first clean the Debatepedia dataset to sample relevant instances from the dataset. Our objective for data sampling here is that the selected samples in the dataset could then be more relevant for query-focused summarization. Afterward, the sampled instances are used for data annotation using ChatGPT. Below we first describe our data sampling technique, followed by our approach of using ChatGPT as the annotator to construct the CQSumDP dataset.

4.1 Cleaned Data Sampling

Our data sampling strategy to use a cleaned version of the dataset for query focused abstractive summarization is as follows:

- We set a minimum threshold of 75 words for the length of each selected document. This is because for the smaller-sized documents, the reference summaries are mainly the overall generic summary of the document where the additional query does not help. By excluding these smaller-sized documents by using a threshold, we can ensure that the reference summaries are more query-focused. Furthermore, setting the threshold at 75 words also helps us to address the noisy scenario in the Debatepedia dataset when the reference summary length is longer than the document length.

- As we demonstrated in Section 3 that many summaries in the Debatepedia dataset are very short (there are many summaries of only 1 word length too), we exclude instances where the length of the summary is shorter than 5 words. This helps us to clean the dataset in a way such that instead of having a dataset with very short answers, we rather propose a dataset consisting of concise but coherent and fluent summaries. This helps us to keep the dataset more relevant to summarization instead of close-ended question answering.

4.2 Using ChatGPT for Data Annotation

As ChatGPT like LLMs has the impressive capability to solve tasks based on the given prompt (Qin et al., 2023; Guo et al., 2023), we manually construct a prompting template that asks ChatGPT to generate the query for a given document-summary pair. Prompt learning is a technique where a machine learning model is trained to complete a task based on the prompted input (Liu et al., 2023; Sanh et al., 2021b). This approach involves presenting the model with a prompt (i.e., a partial input), and the model is then tasked with generating the complete output. Prompt learning has become increasingly popular due to its ability to generate highly accurate results with very little data. It is also highly flexible, as it allows the user to modify the prompt

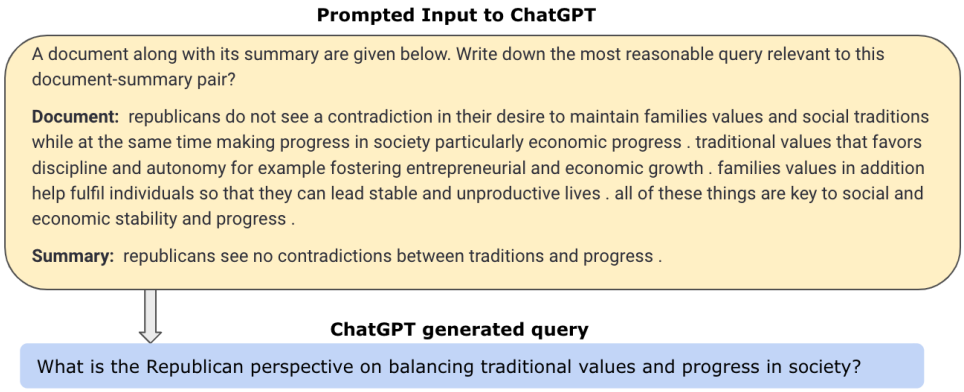

Figure 1: Our Input Prompt to ChatGPT for Query Generation

Split	Total Number of Samples	Avg. Query Length	Avg. Document Length	Avg. Summary Length
Training	5212	11.64	106.82	9.77
Validation	301	11.54	107.22	9.62
Test	401	11.90	104.75	9.77

Table 2: Data distribution on each split (train/valid/test) in our cleaned annotated version of Debatepedia: The CQSumDP Dataset.

to achieve the desired result. We show an example prompt in Figure 1 where ChatGPT is asked to generate a query that is relevant to the given document-summary pair.

The ChatGPT version that we used for data annotation was based on the version that was last released[6] on January 30th. We choose ChatGPT over other text generation models due to its impressive capability of generating high quality responses (Qin et al., 2023; Guo et al., 2023) while being free to use (in contrast to their powerful models in OpenAI that requires the use of paid API subscription). One of the key reasons for ChatGPT to generate human like responses is because it was trained using the reinforcement learning from human feedback technique (Qin et al., 2023; Guo et al., 2023; Ouyang et al., 2022). In this technique, the model generates a response to a user's input, and then humans provide feedback on the quality and appropriateness of the response. This helps the model to generate human like responses while ensuring high accuracy, appropriateness, and fluency. For these reasons, we use ChatGPT for data annotation.

Though prior research has demonstrated that many queries in the Debatepedia dataset have no relevance with the document (Laskar et al., 2022d), there does not have any major issues found on the summaries in the Debatepedia dataset. Thus, we use both the document and the summary as input to ChatGPT since we already cleaned the Debatepedia dataset by removing noisy instances where the summary length is very small or exceeds the document length. While we could ask ChatGPT to generate a query followed by a query-based summary by only giving the document with the input prompt, we did not do so as it has been observed that ChatGPT tends to generate longer summaries (Qin et al., 2023) and so we use both the document and the summary as input to only regenerate the queries in the Debatepedia dataset. This also allows us to use the original gold reference summaries in our proposed CQSumDP dataset without any modification.

A total of 5914 samples were annotated using ChatGPT. After the data annotation is completed, we create the training, validation, and test set based on the split provided by Nema et al. (Nema et al., 2017) for the original version of the Debatepedia dataset[7]. As we construct a cleaned version of the dataset by excluding noisy instances, the number of samples in each split in our cleaned version of the dataset is smaller than the original one. The

[6] https://help.openai.com/en/articles/6825453-chatgpt-release-notes

[7] https://github.com/PrekshaNema25/DiverstiyBasedAttentionMechanism/tree/master/data

Model	Dataset	ROUGE 1	ROUGE 2	ROUGE L
BART-Base	CQSumDP	42.26	22.45	38.84
Pegasus-Base	CQSumDP	36.01	16.30	32.59
T5-Base	CQSumDP	39.95	21.24	36.79
BART-Base	Original Debatepedia	39.97	21.50	36.87
Pegasus-Base	Original Debatepedia	29.70	11.91	26.77
T5-Base	Original Debatepedia	37.68	18.92	34.49

Table 3: Performance of different models trained and evaluated on the respective versions of the Debatepedia dataset.

Model	Training Dataset	Evaluation Dataset	ROUGE 1	ROUGE 2	ROUGE L
BART-Base	CQSumDP	MS-MARCO	44.01	26.95	38.34
Pegasus-Base	CQSumDP	MS-MARCO	50.34	33.07	45.80
T5-Base	CQSumDP	MS-MARCO	48.90	28.66	43.84
BART-Base	Original Debatepedia	MS-MARCO	43.09	23.72	37.90
Pegasus-Base	Original Debatepedia	MS-MARCO	46.94	29.24	42.42
T5-Base	Original Debatepedia	MS-MARCO	47.85	27.89	42.81
BART-Base	MS-MARCO	CQSumDP	28.42	10.30	23.56
BART-Base	MS-MARCO	Original Debatepedia	23.56	7.38	20.88

Table 4: Domain generalization performance of different models trained on respective versions (CQSumDP and Original) of the Debatepedia dataset and evaluated on the MS-MARCO dataset, as well as trained on MS-MARCO and evaluated on the CQSumDP and Original versions of the Debatepedia dataset.

overall statistics of our cleaned, annotated version of the Debatepedia dataset: the CQSumDP dataset is shown in Table 2.

5 Experimental Settings

In this section, we present our experimental settings. Below, we first describe the models we use to evaluate our ChatGPT annotated cleaned version of the Debatepedia dataset, the CQSumDP dataset, followed by our model implementation details. To keep the experimental comparisons fair, we only use the cleaned samples of both versions of the dataset (e.g., 5914 cleaned samples, with 5212, 301, 401 instances in the training, validation, and test sets respectively, as demonstrated in Section 4). From now on, we refer to the version of the Debatepedia dataset that has the original queries but only contains our sampled 5914 instances as *Original Debatepedia*.

5.1 Models

To evaluate the effectiveness of our ChatGPT annotated CQSumDP dataset, we fine-tune some state-of-the-art pre-trained sequence to sequence models (Lewis et al., 2019; Raffel et al., 2019; Zhang et al., 2019; Goodwin et al., 2020). For this purpose, we concatenate the query with the document and give as input to these models to generate the query-focused abstractive summaries as this approach has shown impressive performance in the query-focused abstractive summarization task recently (Laskar et al., 2022d). We describe these models below:

BART (Bidirectional and Auto-Regressive Transformer): BART (Lewis et al., 2019) is a pre-trained sequence-to-sequence model based on the encoder-decoder architecture that was pre-trained on a large amount of diverse text data using the denoising auto-encoding technique to recover the original form of a corrupted document. The pre-training involved various objectives such as rotating the document, permuting sentences, infilling text, masking tokens, and deleting tokens. We use the pre-trained BART model since fine-tuning this model was found to be very effective on a wide range of language generation tasks, including abstractive summarization.

T5 (Text-to-Text Transfer Transformer): The T5 model (Raffel et al., 2019) is a transformer-based model that uses the BERT architecture. Unlike traditional BERT-based models that classify input text into a specific category, the T5 model treats all tasks such as text classification, question answering, neural machine translation, and text summarization as a sequence-to-sequence problem

using various pre-training objectives. After pre-training, the model is fine-tuned to generate the output for a given input sequence in the required task, leading to impressive performance gain on many downstream summarization datasets.

Pegasus (Pre-training with Extracted Gap-sentences for Abstractive Summarization): Pegasus (Zhang et al., 2019) is a transformer-based pre-trained encoder-decoder model for abstractive summarization. Its pre-training objective involves generating summary like text from an input document. To achieve this, the PEGASUS model first selects and masks some sentences from the input document(s). It then concatenates these selected sentences to create a pseudo-summary. The model uses different approaches to select these sentences, such as randomly selecting a certain number of sentences, selecting the first few sentences, or computing the ROUGE-1 score between each sentence and the rest of the document to choose the top-scoring sentences. This pseudo-summary is then used for self-supervised learning. By pre-training on large datasets using this approach, the model achieves impressive fine-tuning performance on downstream summarization datasets.

5.2 Implementation

We use the HuggingFace[8] (Wolf et al., 2019)library to implement the baseline models for performance evaluation. Similar to the prior work, we concatenated the query with the document to give as input to the pre-trained baselines (i.e., BART, Pegasus, T5). The pre-trained model is then fine-tuned using 4 NVIDIA V100 GPUs. The training batch size for BART was set to 16, while it was set to 4 for Pegasus and T5. The other hyperparameters were similar for all models, with the learning rate being set to $2e-3$ and the maximum input (i.e., the concatenated query and document) sequence length being 150 tokens. The minimum and the maximum target (i.e., the generated summary) sequence lengths were 5 and 25, respectively. A total of 10 epochs were run to fine-tune the pre-trained summarization models. We computed the ROUGE (Lin, 2004) scores in terms of ROUGE-1, ROUGE-2, and ROUGE-L using the *Evaluate*[9] library to compare the performance of different models on the respective test set.

[8] https://huggingface.co/

[9] https://huggingface.co/spaces/evaluate-metric/rouge

6 Results & Discussions

We conduct a series of experiments to evaluate the performance of strong baseline models in our proposed cleaned annotated version of Debatepedia: the CQSumDP dataset. In this section, we present our experimental findings.

6.1 Effectiveness of ChatGPT Generated Queries

To investigate the effectiveness of our CQSumDP dataset that leverages ChatGPT to generate the queries, we compare the performance of BART, Pegasus, and T5 models on both the CQSumDP and the Original Debatepedia datasets (results are given in Table 3). We use the Base versions of these models from HuggingFace (Wolf et al., 2019), and trained and evaluated on the respective datasets.

From Table 3, we find that all three models perform better in the CQSumDP dataset in comparison to their performance on the Debatepedia dataset. This gives a strong indication that the queries generated by ChatGPT are more helpful to improve the model performance. While comparing the performance between different models, we found that BART outperforms the other two models on both datasets in all three ROUGE metrics. More specifically, in the CQSumDP dataset, BART achieves the highest ROUGE-1 (42.26), ROUGE-2 (22.45), and ROUGE-L (38.84) scores. Though in the Original Debatepedia dataset, BART also outperforms other models by achieving ROUGE-1, 2, and L scores of 39.97, 21.50, and 36.87, respectively; its performance on the Original Debatepedia is much lower than its performance on the CQSumDP dataset.

Our experimental results show the effectiveness of our proposed CQSumDP dataset that helps all these models to obtain better ROUGE scores than their counterparts on the Original Debatepedia dataset. The poor performance of these models on the Original Debatepedia dataset compared to the CQSumDP dataset further demonstrates the limitations in terms of query relevance in the Original Debatepedia.

6.2 Generalization Capability

In the previous section, we find that all the baseline models fine-tuned on our CQSumDP dataset perform better than their counterparts that are fine-tuned on the Original Debatepedia dataset. In this section, to further study the relevance of the ChatGPT generated queries in our proposed CQSumDP

Model	ROUGE-1	ROUGE-2	ROUGE-L
BART-Large	51.66	33.96	49.03
BART-Base	42.26	22.45	38.84

Table 5: Performance comparisons based on model size between BART-Large and BART-Base on CQSumDP.

Model	ROUGE-1	ROUGE-2	ROUGE-L
BART-Large	51.66	33.96	49.03
without query incorporation	46.45	29.92	44.11

Table 6: Ablation test results after removing the query relevance in the CQSumDP dataset.

dataset, we evaluate the performance based on domain generalization. In this regard, we use the similar setting of Laskar et al., (Laskar et al., 2022d) where they used the QA-NLG version of the MS-MARCO dataset (Wang et al., 2018) to fine-tune their query-focused summarization model for abstractive answer generation and then evaluate on Debatepedia. We also use the MS-MARCO dataset for our analysis based on the following two scenarios:

- **Training: MS-MARCO, Evaluation: Debatepedia:** In this scenario, we trained the baseline models on the training set of MS-MARCO (153725 samples) and evaluated on the respective versions of the Debatepedia dataset (CQSumDP and Original Debatepedia).

- **Training: Debatepedia, Evaluation: MS-MARCO:** In this scenario, we do the opposite, as we trained the baseline models on the respective versions of Debatepedia and evaluated on the development set of MS-MARCO (12467 samples).

We show our results in Table 4 and observe that the domain generalization performance in both scenarios: (i) while using Debatepedia for training to evaluate on MS-MARCO, as well as (ii) while using MS-MARCO as the training data for evaluation on Debatepedia, the performance is better when the CQSumDP version of the Debatepedia dataset is used in comparison to the scenario when the Original Debatepedia is used. These findings further establish the effectiveness of using ChatGPT generated queries for the query-focused summarization task in the Debatepedia dataset.

6.3 Performance Based on Model Scaling

So far, in our prior experiments, we utilize the Base version of each model and investigate the effectiveness of our proposed CQSumDP dataset. Though smaller models are preferred over larger models in real-world industrial scenarios where computing resources are limited (Laskar et al., 2022b,a), in this section, to set a benchmark performance in our proposed CQSumDP dataset, we investigate how much performance gain we can achieve via scaling to a larger model. For this purpose, we select the best performing BART model (Lewis et al., 2019) and compare its performance between its Base and Large versions in our dataset. From our experimental results given in Table 5, we observe that the ROUGE score is improved by a large margin (on average an improvement of 10.37 out of all three ROUGE metrics) when the BART-Large model is used. This indicates that the utilization of the ChatGPT generated queries in the CQSumDP dataset also helps the larger summarization models to understand the query representation better, leading to an improved ROUGE score.

6.4 Ablation Tests

It was recently found that even without incorporating query relevance, the summarization models could achieve performance on the Debatepedia dataset almost similar to what could have been achieved via incorporating query relevance (Laskar et al., 2022d). While analyzing the Debatepedia dataset, we observe that this happens mostly on instances where the document size is quite small. As we already cleaned the Debatepedia dataset by removing such instances (e.g., short documents or summaries), in this section, we conduct ablation studies to investigate the importance of query relevance in the cleaned version of the dataset. For this purpose, we remove the query relevance while giv-

Model	Evaluation Dataset	ROUGE 1	ROUGE 2	ROUGE L
Pre-trained BART-Large	CQSumDP	26.86	9.46	21.70
Pre-trained BART-Large	Original Debatepedia	21.60	6.04	18.52

Table 7: Zero-Shot Learning Performance of different models on the respective evaluation sets of Debatepedia.

#	Original Query	ChatGPT Query	Source Document	Gold Summary
1.	military :	What actions did the government take to improve the situation for U.S. troops and veterans?	provided better body armor to our troops . provided the department of veterans affairs (va) with more than $ # . # billion to improve services to america s veterans . ended media blackout on war casualties and the return of fallen soldiers to dover afb . announced creation of a joint virtual lifetime electronic record for members of the u.s. armed forces to improve quality of medical care . ended the previous stop-loss policy that kept soldiers in iraq/afghanistan longer than their enlistment date . signed the veterans health care budget reform and transparency act authorizing advance appropriations for the department of veterans affairs by providing two-fiscal year budget authority thus enabling better medical care for veterans . endorsed by the american legion american veterans blinded veter ... ans association	improved services benefits and respect for troops .
2.	we economy : has wto benefited the economy of the united states ?	Has NAFTA caused job losses in the U.S?	" nafta and job losses " . cyril morong (PhD) the wall street journal may # # - " did nafta cause the u.s. to lose so many jobs [citing figures provided in the range of # million and # #] especially high-paying manufacture jobs ? probably not . i say probably since causality in any social science (economics included) is difficult to prove since so many factors change so quickly in the real world . but if many high-paying manufacture jobs were lost it took many years until after nafta went into effect before they were ... but what about manufacture jobs ? we had just about # million in # . it actually rose to # . # million in # and was at # . # in #.	nafta has decreased the number of american job
3.	entrepreneurs: does an mba help entrepreneurs ?	Is an MBA necessary for product managers?	christopher cummings . " is an mba necessary for product managers ? " product management meet pop culture . february # # : " hindsight . looking back the brass tacks of my mba experience were about the basics of management economics and business strategy . could that have been picked up on the job ? maybe . [...] however the more important throughline of the experience relates to critical thinking perspective and learning when to lead and when to follow . [...] on the job especially as a young pm it can be easy to lose perspective to miss the forest for the trees . at the time i was definitely into the plate-spinning the go-go-go the tactics and day-to-day . no time to think ; just keep moving . [...] the mba experience	mba teach strategy plan not just tactics

Table 8: Comparisons between the original queries and the ChatGPT generated queries in some samples of the Debatepedia dataset. Note that the personally identifiable information in this dataset is anonymized with the # token.

ing the input to the best performing BART-Large model and investigate the effect of removing the query in our proposed dataset. We show our results in Table 6. We find from the table that the performance is dropped by a huge margin when the query is removed from the input text, demonstrating the importance of the query in our proposed CQSumDP dataset.

6.5 Zero-Shot Learning Performance

In recent times, the zero-shot evaluation of large pre-trained language models on text generation tasks, such as abstractive summarization has been on the rise (Brown et al., 2020; Qin et al., 2023; Guo et al., 2023). To establish a benchmark in our proposed dataset, we also conduct a zero-shot evaluation of the best performing BART-Large model in both the CQSumDP and the Original Debatepedia datasets. To do so, we combine the query with the document and give as input to the pre-trained BART-Large model. We observe from Table 7 that in terms of zero-shot evaluation, the pre-trained BART-Large model evaluated on our dataset performs better than its performance on the Original Debatepedia, further establishing that the utilization of ChatGPT generated queries in CQSumDP is more helpful than the original queries in the Debatepedia dataset.

6.6 Qualitative Analysis of the Annotated Data

In this section, we do some qualitative analyses between the queries in the Original Debatepedia dataset as well as the queries generated using ChatGPT in our proposed CQSumDP version of the Debatepedia dataset. For our analysis, we collect a set of 3 samples from this dataset and present them in Table 8. While comparing between the queries in the first example in the table, we find that the original query is just one word length and very

ambiguous, while the ChatGPT generated query is more descriptive and more relevant to both the document and the summary. For the second example, we find that even though the original query is descriptive, it does not have any relevance to the generated summary. Whereas the ChatGPT generated query is very relevant to both the document and the summary. For the third example, we find that the original query is related to "entrepreneurs". However, the document is about "product managers", not "entrepreneurs". Meanwhile, the ChatGPT generated query is also very relevant to the document. This analysis further demonstrates the relevance of our ChatGPT generated query in comparison to the original query in Debatepedia.

6.7 Cost Efficiency Analysis

Recently, it was shown that using the GPT-3 (Brown et al., 2020) model could significantly reduce the labeling cost without sacrificing the model's performance much, making it possible to train models on larger datasets without the need for extensive manual labeling (Wang et al., 2021; Ding et al., 2022). However, to use GPT-3, it requires the use of its API[10], which is not free. On the contrary, ChatGPT is free to use. Meanwhile, we observe that generating the query in the Debatepedia dataset was also quite fast, as we observe that we could generate the queries for about 4 samples on average per minute while using ChatGPT for data annotation. This is also quite fast than giving the data for human annotation, as the human not only needs to read the document and the summary, but also needs some time to think about what could be the most effective query for the given document-summary pairs. Thus, in terms of both cost and time, it is more efficient to use ChatGPT for data annotation.

7 Conclusions and Future Work

In this paper, we presented a methodology for cleaning the Debatepedia dataset to make it suitable for query-focused abstractive summarization. We removed the noise from the dataset to construct a cleaned version of the dataset while using ChatGPT's language generation capabilities to address the limitations of the queries in this dataset. Our approach results in a cleaner version of Debatepedia that is found to be very effective for training and evaluating query-focused summarization models via outperforming the original dataset in terms of query relevance and summary generation quality. This indicates that our cleaning approach is effective in improving the dataset's quality for research in summarization.

In the future, we will explore if the chain of thought prompts (Wei et al., 2022) with ChatGPT leads to better query generation. We will also explore the performance of fine-tuning other pre-trained models on our proposed dataset (Sanh et al., 2021c; Muennighoff et al., 2022; Chowdhery et al., 2022). In addition, we will investigate the potential of using ChatGPT as the annotator for other tasks in Information Retrieval (Lin et al., 2021; Laskar et al., 2020c, 2022c,a; Huang and Hu, 2009; Huang et al., 2005; Liu et al., 2007) to assess its generalizability. Finally, we will release our annotated version of the Debatepedia: the proposed CQSumDP dataset to encourage further research in the query-focused abstractive summarization task.

Acknowledgements

We would like to thank OpenAI for making ChatGPT freely available which helps us to use it for data annotation. This research is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the York Research Chairs (YRC) program.

8 Limitations

ChatGPT itself is continuously upgraded by OpenAI. Meanwhile, the ChatGPT-generated responses are quite random. Thus, it may not be possible to reproduce the same queries using ChatGPT. Nonetheless, this also mimics the real-world scenario as different human annotators may write different queries (e.g., in many text summarization datasets, there can have multiple gold reference summaries written by different human annotators). ChatGPT-generated responses are also random, as it may generate different responses for the same input at different times. However, similar to the work of Guo et al. (2023), we also notice that this difference is very small and so we also generate only one query for each example. Though a new version of ChatGPT called GPT-4 (**?**) has been published which may generate more powerful queries, in this work, we did not utilize GPT-4 as it is only accessible for the ChatGPT plus subscribers. Also, it is more expensive than the original ChatGPT while being significantly slower. Nonetheless, fu-

[10] https://platform.openai.com/docs/models

ture work may compare with other more powerful LLMs (including GPT-4) for data annotation.

9 Ethics Statement

This paper does not leverage any 3rd-party to generate the ChatGPT responses and so no additional compensation was needed. Since this paper only utilizes ChatGPT to generate the queries for the given document-summary pairs, it does not lead to any unwanted biases or ethical concerns. However, all the responses generated by ChatGPT are still manually checked by the authors to ensure that the ChatGPT-generated queries in the cleaned version of the dataset do not pose any ethical concerns or unwanted biases as well as do not contain any nonsensical queries due to hallucination. Only a publicly available academic dataset is used that did not require any licensing. Thus, no personally identifiable information has been used while utilizing ChatGPT to fix the queries in the Debatepedia dataset.

References

Deen Mohammad Abdullah and Yllias Chali. 2020. Towards generating query to perform query focused abstractive summarization using pre-trained model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 80–85.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2019. Joint learning of answer selection and answer summary generation in community question answering. *arXiv preprint arXiv:1911.09801*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.*

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 961–964.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*.

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Flight of the pegasus? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5640–5646.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Xiangji Huang and Qinmin Hu. 2009. A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314.

Xiangji Huang, Ming Zhong, and Luo Si. 2005. York University at TREC 2005: Genomics track. In *Proceedings of the Fourteenth Text REtrieval Conference, TREC*, pages 56–59.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: Beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Md Tahmid Rahman Laskar, Cheng Chen, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan TN, and Simon Corston-Oliver. 2022a. An auto encoder-based dimensionality reduction technique for efficient entity linking in business phone conversations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3363–3367.

Md Tahmid Rahman Laskar, Cheng Chen, Aliaksandr Martsinovich, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan Tn, and Simon Corston-Oliver. 2022b. Blink with elasticsearch for efficient entity linking in business conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 344–352.

Md Tahmid Rahman Laskar, Cheng Chen, Aliaksandr Martsinovich, Jonathan Johnston, Xue-Yong Fu, Shashi Bhushan Tn, and Simon Corston-Oliver. 2022c. BLINK with Elasticsearch for efficient entity linking in business conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 344–352, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Huang. 2020a. Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Canadian Conference on Artificial Intelligence*, pages 342–348. Springer.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022d. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

Md Tahmid Rahman Laskar, Enamul Hoque, and Xiangji Huang. 2020b. WSL-DS: Weakly supervised learning with distant supervision for query focused multi-document abstractive summarization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5647–5654.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020c. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5505–5514.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 607–614.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3721–3731.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Preksha Nema, Mitesh M Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Haggai Roitman, Guy Feigenblat, Doron Cohen, Odellia Boni, and David Konopnicki. 2020. Unsupervised dual-cascade learning with pseudo-feedback distillation for query-focused extractive summarization. In *Proceedings of The Web Conference 2020*, pages 2577–2584.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021a. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021b. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021c. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7(1):1–9.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. 2020. Caire-covid: a question answering and query-focused multi-document summarization system for covid-19 scholarly information management. *arXiv preprint arXiv:2005.03975*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2022. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jesse Vig, Alexander R Fabbri, Wojciech Kryściński, Chien-Sheng Wu, and Wenhao Liu. 2021. Exploring neural models for query-focused summarization. *arXiv preprint arXiv:2112.07637*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018. Multi-passage machine reading comprehension with cross-passage answer verification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1918–1927.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645.

Yumo Xu and Mirella Lapata. 2021. Text summarization with latent queries. *arXiv preprint arXiv:2106.00104*.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Jin-Ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53(2):297–336.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, et al. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.