David vs. Goliath: Can small models leverage LLMs for summarization?

Anonymous ACL submission

Abstract

Recent studies indicate a preference for summaries generated using large language models (LLMs) over those using classical models, highlighting a performance discrepancy. This study explores strategies to narrow the gap between the summaries generated through these two models. To address this, we introduce a novel framework that uses LLM-generated summaries to train classical models, adopting a two-stage training approach to enhance their summary quality. Although classical models are relatively smaller in size than LLMs, through automatic metrics and human evaluations, we can demonstrate that the performances of classical models, trained using LLM-generated references can catch up with LLMs. Our findings create a simple yet potential way to improve classical summarization models by leveraging LLMs. Additionally, we contribute a new dataset **GXSum**[1], enabling further research and promoting development progress in this subject.

1 Introduction

Text summarization plays a pivotal role in the field of natural language processing by condensing articles into concise versions that capture the main information. With the rapid development of deep learning, automatic text summarization systems have made significant progress. (Nallapati et al., 2016a; Vaswani et al., 2017; Li et al., 2018; Shi et al., 2021). More recently, large language models (LLMs) have revolutionized the field of natural language processing. These models exhibit remarkable results in summarization accuracy, particularly under zero-shot and few-shot fine-tuning scenarios (Wang et al., 2023; Basyal and Sanghvi, 2023; Ahmed and Devanbu, 2023). Unlike classical models, LLMs leverage reinforcement learning from human feedback (RLHF) (Kirk et al., 2023), fine-tuning their outputs to align more closely with human preferences, thereby widening the performance gap with classical models (Wang et al., 2023; Zhang et al., 2024; Fabbri et al., 2021). Some studies even indicate that humans might prefer LLM-generated summaries to those written (or selected) by humans (Liu et al., 2023b,a).

Sweeping over previous research on text summarization, most studies mainly concentrated on developing novel model architectures (Dou et al., 2021; Wang et al., 2022a; Liu et al., 2022) or training method (Stiennon et al., 2020). These efforts improve performance on specific benchmarks, yet they often increase model complexity or compromise training efficiency. However, these efforts still do not bridge the performance gap with LLMs.

Knowledge distillation is a simple and straightforward way to transfer model capabilities from one model to another. To move beyond LLMs in a simple and cost-effective way, we present a two-stage training framework that is expected to allow classical summarization models to rival the performance of LLMs based on the fundamental philosophy of knowledge distillation in this study. More specifically, in the first stage, we leverage LLMs to generate summaries and form a new dataset. Next, we train classical models referring to the new ground truths with the traditional maximum likelihood objective. By doing so, the classical model is expected to not only inherit the advantages of LLMs but also retain the abilities of original designs, delivering better results than LLMs.

In sum, our key contributions are at least threefold. First, we propose a simple yet efficient framework to enhance the performance of classical models and catch up with LLMs. Second, a series of experiments were used to show that significant performance gains are achievable even with limited data for fine-tuning. Of course, as always, more data yields better results. Third, a new dataset **GXSum** is released to facilitate further research,

[*] Equal contribution.

[1] https://github.com/anonymous

perform fair comparisons, make results producible, and promote research progress in the line of research.

2 Related Work

Previous research has demonstrated the exceptional proficiency of LLMs in generating summaries, outperforming classical models in both automated evaluation metrics and human assessments. Additionally, summaries generated using LLMs, especially in the news domain, have been shown to be at par with, or even superior to, those crafted by humans. These results reveal significant potential for LLMs on the text summarization task (Victor et al., 2022; Wang et al., 2022b; Goyal et al., 2022). Some studies further emphasize that the field of summarization is undergoing significant changes, suggesting a pivotal moment in summarization research. A thought-provoking question is whether those human-generated ground truths bound the performances of classical summarization models (Pu et al., 2023; Zhang et al., 2024).

The feasibility of using LLMs for generating source data has been extensively explored. Some research has introduced methods for distilling LLMs and employing them in data augmentation tasks (Wang et al., 2021; Ding et al., 2023; Kang et al., 2023). Specifically, these methods focus on extracting the most relevant information from LLMs to enrich training datasets, thereby enhancing model performance without the need for extensive computational resources. Notably, a series of studies have demonstrated the use of LLMs to generate both final answers and task-related descriptions, which aid in training smaller models for reasoning tasks (Li et al., 2022; Shridhar et al., 2023; Hsieh et al., 2023). In the realm of text summarization, Wang et al. (2021) have used GPT-3 (Brown et al., 2020) to generate reference summaries. Concurrently, Gekhman et al. (2023) proposed the use of LLMs for annotating summary factual consistency (Maynez et al., 2020), facilitating the training of models to evaluate factual consistency. Moreover, Liu et al. (2023c) have explored further fine-tuning of news summaries generated by the GPT series for the summarization domain.

Therefore, in this paper, we expand the dataset and thoroughly analyze the differences between LLMs and human summarization. In the subsequent research, we will further train the summaries generated using LLMs, aiming to redefine the role of LLMs in summarization tasks.

3 LLM-Guided Summarization

3.1 Models

In this study, we selected the most advanced ChatGPT[2] provided by OpenAI as an example. To minimize the randomness of generated results, we set the `temperature` parameter of the model to 0, whereas other parameters are at their default values to ensure stability and reproducibility of the experimental results.

For a comprehensive analysis, BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and BRIO (Liu et al., 2022) were chosen as the basic classic summarization models for our experiments. These models have been proven in previous research to possess excellent text summarization capabilities, each representing various research directions in the field of summarization. The pre-trained models of BART and PEGASUS are sourced from the Transformers Library (Wolf et al., 2020), whereas the weights for BRIO are obtained from the GitHub repository of the original paper.

3.2 Human Referenced Datasets

In this study, we adopted two key news summarization datasets that are widely used in the research of summarization models and the evaluation of large language model performance: the Extreme Summarization Dataset (abbreviated as XSum) (Narayan et al., 2018)[3] and the CNN / DailyMail News Summarization Dataset (abbreviated as CNNDM)[4] (Nallapati et al., 2016b). The XSum dataset is comprised of press releases from the British Broadcasting Corporation, whereas the CNNDM dataset compiles news articles from the Cable News Network (CNN) and the Daily Mail. Notably, these two datasets differ significantly in their nature. Compared to CNNDM, the summary reference texts in XSum mostly contain only one to two sentences, posing a significant challenge for summarization models to refine and extract core information for the summary. Table 1 shows the ROUGE scores (cf. section 4.2) of classic models on the XSum and CNNDM datasets.

[2] GPT-4-Turbo (gpt-4-1106-review) https://platform.openai.com/docs/models/overview

[3] https://github.com/EdinburghNLP/XSum

[4] https://cs.nyu.edu/~kcho/DMQA/

	XSum			**CNNDM**		
Models	R-1	R-2	R-L	R-1	R-2	R-L
BART	45.14	22.27	37.25	44.16	21.28	40.90
PEGASUS	47.21	24.56	39.25	44.17	21.47	41.11
BRIO	49.07	25.59	40.40	47.78	23.55	44.57

Table 1: BART, PEGASUS, and BRIO's ROUGE scores on the XSum and CNNDM datasets.

3.3 LLM Referenced Dataset

As one of the core objectives of our research, we created a dataset comprising summaries generated by LLMs to serve as reference summaries. This dataset is based on XSum and CNNDM, maintaining the format of the original datasets. To leverage the ChatGPT API for generating high-quality summaries, we have meticulously designed a prompt template that specifically emphasizes the role of ChatGPT as a summary writer. Additionally, to better control the summary length, we included a description of the length limit as a soft constraint in the prompt and set the API `max_tokens` parameter as a hard constraint. The detailed design of the prompt is presented in Appendix A.1. For the source text, we designated the `document` from XSum and the `article` from CNNDM as the variables. During the summary generation process, the length restriction was set to ensure that the difference in lengths between the newly generated summaries and the original reference summaries remained within a range of plus or minus five tokens. We provide an example of our summary generation process in Appendix A.2.

3.4 Implementation Details

3.4.1 Data Processing

We extracted a sample comprising 20,000 data points from the training set and 1,100 data points from the validation set. These samples were subjected to the LLM summarization workflow to produce reference summaries. This subset was designated as the ***Small*** variant. In contrast, the test set underwent comprehensive processing to guarantee a robust and reliable evaluation. Data processing was conducted on both the XSum and CNNDM datasets to ensure uniformity and accuracy in our analyses.

3.4.2 Training Details

The initiation of training for each model leveraged checkpoints that had been previously fine-tuned on the benchmarked XSum and CNNDM datasets. These fine-tuned checkpoints used for BART [5], PEGASUS [6] and BRIO [7] were obtained from the Huggingface library. For optimization, the AdamW optimizer was employed, incorporating a weight decay of 0.01 and an initial learning rate of 0.00002. A linear learning rate scheduler was applied without any warm-up steps. Model performance evaluation on the validation set informed the selection of checkpoints, whereas performance metrics on the test set were documented and reported.

3.5 Evaluation Methods

To validate the performance of our model, we use two primary evaluation methods: human validation and automatic metrics. Initially, human validation gauges the summaries' quality from readers' viewpoints. Automatic metrics are used to determine whether the fine-tuning process is functioning properly and toward the training objectives.

3.5.1 Human Evaluation Protocol

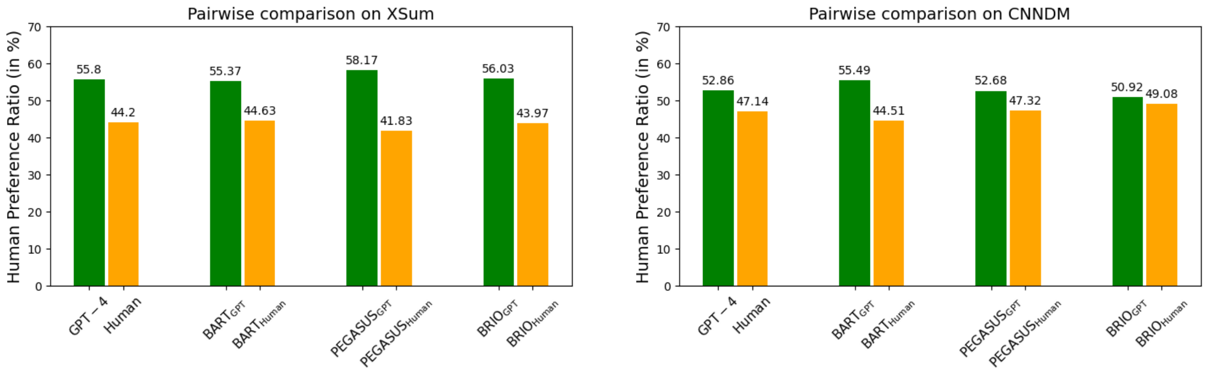
As the main evaluation methods of this study, we adopted three common forms of human validation, including the Likert scale scoring, pairwise comparison, and multiple candidate ranking.

The Likert scale scoring is the most used method in human validation assessments. The evaluation process involves presenting a source text and its corresponding generated summary, where human annotators are required to score the summary on several aspects of performance. In this research, we defined five distinct aspects for evaluation: relevance, consistency, fluency, coherence, and infor-

[5] https://huggingface.co/facebook/bart-large-xsum and https://huggingface.co/facebook/bart-large-cnn

[6] https://huggingface.co/google/pegasus-xsum and https://huggingface.co/google/pegasus-cnn_dailymail

[7] https://huggingface.co/Yale-LILY/brio-xsum-cased and https://huggingface.co/Yale-LILY/brio-cnndm-cased



(a)

(b)

Figure 1: Pairwise Comparison on XSum and CNNDM

mativeness. Detailed guidelines for these metrics are elaborated in Appendix B.1. Through these metrics, human annotators can more comprehensively score the overall quality of summaries. The scoring range is set from 1 (worst) to 5 (best).

Pairwise comparison is a human validation evaluation method based on relative comparison. Given a source text and two summaries generated by different models, assessors are asked to select the one with the better quality.

Multiple candidate rating is an advanced and complex variation of the pairwise comparison method. Assessors are compelled to examine a set of summaries for a given source text and assign a unique rating to each, reflecting the overall quality of each summary. Therefore, the method facilitates a thorough evaluation of the performance variations across various summarization models. Within our experiment, we established a rating scale from 1 (lowest quality) to 5 (highest quality).

3.5.2 Automatic Evaluation Metrics

We adopted Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004) as our automatic evaluation metric for summarization effectiveness. ROUGE is crucial in performing summarization research, serving as a standard for comparing the similarity and quality between computer-generated and human-crafted reference summaries. This study employs three ROUGE variants: ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L). ROUGE-1 assesses unigram similarity to gauge informational content. ROUGE-2 evaluates bigram similarity for fluency. ROUGE-L focuses on the longest common subsequence to determine core content extraction.

4 Experiment Result and Analysis

4.1 Human preference

The collection of human annotations contains evaluations of summaries generated by models that were fine-tuned on the ***Small*** dataset. These evaluations were obtained through a combination of crowd-sourced contributors and expert judgments.

4.2 Crowd-Sourced Annotations

We gathered annotations via Amazon Mechanical Turk (MTurk) for 1,000 articles from the XSum and CNNDM test sets. Details of the recruitment process are in Appendix B.2. We compared models fine-tuned on the original datasets with those fine-tuned on the LLM-reference dataset. Each summary was evaluated by three annotators using Likert scale scoring and pairwise comparison methods (Section 3.5.1).

Figures 1b and 1a show the crowd-sourced winning rates from pairwise comparisons. Systems trained using human references are denoted with ***Human***, while those using GPT-4 references are marked with ***GPT***. Key observations:

(1) GPT-4 generated summaries were preferred over human-written ones for both XSum and CNNDM tasks, supporting hypotheses from related works (Goyal et al., 2022; Liu et al., 2023c; Pu et al., 2023).

(2) Models trained on GPT-4 references consistently outperformed those trained on human references, demonstrating the benefits of high-quality, AI-generated references in supervised training.

(3) The performance advantage was less pronounced for CNNDM compared to XSum.

Table 2 shows the Likert scale scoring results. Our analysis revealed:

Dataset	System	Relevance	Consistency	Fluency	Coherence	Informativeness
	Human Base	0%	0%	0%	0%	0%
XSum	GPT-4	**+13.8%**	**+13.2%**	**+9.3%**	**+7.5%**	**+3.6%**
	BART$_{GPT}$	**+17%**	**+15.5%**	**+10.9%**	**+11.3%**	**+4.2%**
	PEGASUS$_{GPT}$	**+18.3%**	**+15.4%**	**+14.5%**	**+16.5%**	**+7.4%**
	BRIO$_{GPT}$	**+11%**	**+8.3%**	**+9%**	**+7%**	**+3.3%**
CNNDM	GPT-4	**+3.58%**	**+1.6%**	**+5.6%**	**+1.2%**	**-0.2%**
	BART$_{GPT}$	**+0.2%**	**+0.7%**	**+1.4%**	**+1.4%**	**+0.9%**
	PEGASUS$_{GPT}$	**-1.1%**	**+3.1%**	**+1.5%**	**+1.8%**	**+1.4%**
	BRIO$_{GPT}$	**-1.9%**	**+2.9%**	**+0.7%**	**+0.9%**	**-0.5%**

Table 2: Evaluation through Crowd-Sourced Likert Scale Scoring, which models referenced by humans serve as the baseline for comparison (default as 0%). The report highlights the percentage difference in occurrences where one system is adjudged to *outperform* the other. For instance, GPT-4 exceeds human writers in Relevance by 13.8% on the XSum dataset. In case of a tie, both systems are recognized as winners.

(1) For XSum, GPT-4 referenced models outperformed across all metrics, with the most significant improvement in summary relevance. Informativeness remained the weakest point due to the dataset's requirement for highly abstract, single-sentence summaries.

(2) For CNNDM, GPT-4 referenced summaries still outperformed human-generated ones, but the margin was narrower (often within 1-2%). This is likely due to the dataset's approach of collecting human-written summary bullets, which tend to be more extractive and closely mirror the original content.

While these results validate our LLM-guided training approach, we acknowledge potential reliability concerns due to variability in nonexpert judgments (Callison-Burch and Dredze, 2010; Goyal et al., 2022; Zhang et al., 2024). To address this, we conducted additional analyses with expert reviewers for more dependable evaluations.

4.2.1 Expert Annotations

To ensure the rigor of expert analysis, we established specific criteria for the selection of annotators, focusing on those with a requisite level of expertise. We collected annotations for a sample of 100 articles from the XSum and CNNDM test sets. The evaluation of each summary was entrusted to three distinct expert annotators who applied the Multiple Candidates Rating Methods as delineated in Section 3.5.1. Additionally, annotators were required to provide reviews of their annotations, enabling verification of results. The candidates the position of an expert annotator are hired from the Upwork platform. The detailed recruitment setting is described in Appendix B.3.

Figures 2b and 2a illustrate the rating distributions (1-5) for each system according to expert evaluations. The analysis yields two key insights:

(1) Expert raters show a clear preference for summaries generated using GPT-4 and GPT-4-assisted systems over those written by humans. This supports our hypothesis based on crowd-sourced annotations, confirming the ability of our system to produce summaries more aligned with human preferences.

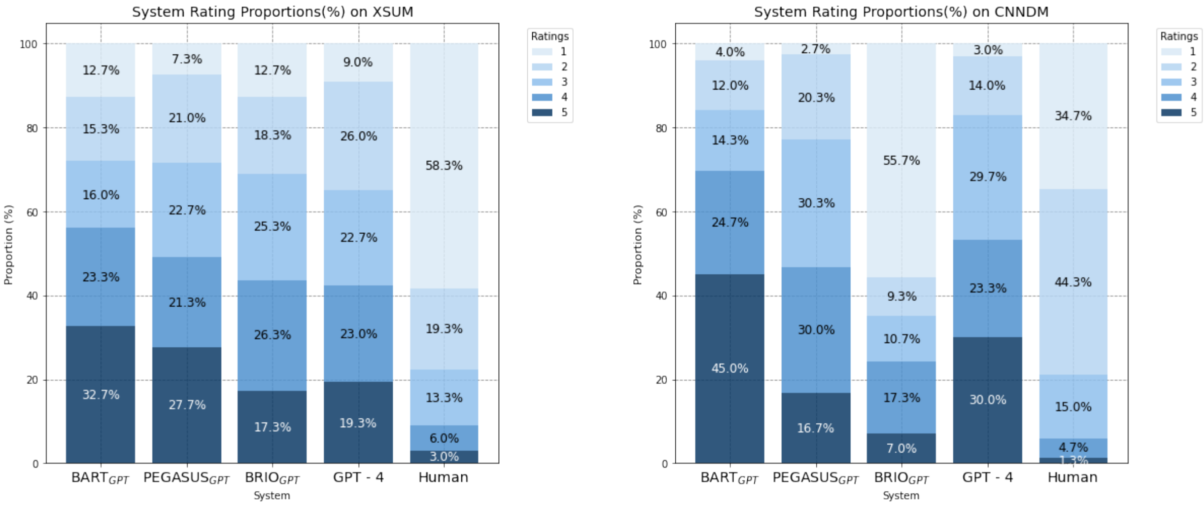
(2) Notably, **models trained using GPT-4 references achieve, and sometimes surpass, the performance of GPT-4 in expert assessments, reaching a 68% inter-annotator agreement**. This indicates that using our training methodology, smaller models can attain the efficacy of large language models.

(3) In both datasets, BART$_{GPT}$ performs better than GPT-4, with its proportion in CNNDM reaching as high as 45%. However, BRIO$_{GPT}$'s performance in CNNDM is closer to that of humans.

XSum	CNNDM
0.3187	0.4377

Table 3: Ranking Pearson correlation coefficient by three raters.

Based on the observed results, we further calculated the Pearson correlation coefficient (Pearson, 1907) for expert rankings, as shown in Table 3. The



(a)

(b)

Figure 2: Rating proportions in XSum and CNNDM

XSum		CNNDM	
Gwet's AC1	PA	Gwet's AC1	PA
0.2252	41%	0.3773	52%

Table 4: Gwet's coefficient and Percentage agreement of BART$_{GPT}$ better than GPT-4 by three raters.

XSum		CNNDM	
Gwet's AC1	PA	Gwet's AC1	PA
0.4081	48%	0.5940	60%

Table 5: Gwet's coefficient and Percentage agreement of BART$_{GPT}$ and PEGASUS$_{GPT}$ being better than human by three raters.

correlation coefficient for XSum is 0.3187, and for CNNDM is 0.4377, indicating a positive correlation, meaning that the rankings by experts tended to be consistent. We also analyzed the consistency and agreement among the three experts who unanimously considered BART$_{GPT}$ better than GPT-4. Based on the results in Figure 2, we observed no significant difference between BRIO$_{GPT}$ and human. Therefore, we compared the consistency and agreement among the three experts who unanimously considered BART$_{GPT}$ and PEGASUS$_{GPT}$ better than human. We used Gwet's Coefficient (Gwet's AC1) (Gwet, 2008) and Percentage Agreement (PA) for calculations, with results shown in Tables 4 and 5, respectively. For the consistency and agreement in considering BART$_{GPT}$ better than GPT-4, the consistency for both datasets simultaneously is 0.2252 and 0.3773, respectively, which is considered Fair agreement. The agreement percentages are 41% and 52%, respectively. For the consistency in considering BART$_{GPT}$ and PEGASUS$_{GPT}$ better than human, the values are 0.4081 and 0.5940, respectively, which is considered Moderate agreement, with agreement percentages of 48% and 60%. Therefore, both in terms of ranking proportions and consistency and agreement, these results demonstrate that by incorporating GPT-4 generated summaries into smaller models through simple knowledge distillation, we have achieved summarization capabilities comparable to GPT-4.

4.2.2 Brief

The results of the multi-summary ranking are quite remarkable. We employed a teacher forcing training method, and theoretically, the model should not surpass the teacher's performance. However, after using GPT-4 generated summaries, it outperformed GPT-4 in expert rankings. This conclusion further enhances the significance of this research. Additionally, although CNNDM showed higher consistency and agreement values than XSum, XSum performed better in overall pairwise comparisons and rankings. Therefore, this paper proposes GX-Sum, a news summarization dataset consisting of numerous GPT-4-generated summaries.

4.3 Automatic Metric

Next, we compare various summary generation models on the XSum and CNNDM datasets, us-

Reference	Hypothesis	XSum R-1	XSum R-2	XSum R-L	CNNDM R-1	CNNDM R-2	CNNDM R-L
GPT-4	Human	24.95	5.64	18.59	36.80	10.89	31.91
	BART$_{GPT}$	45.36	19.59	36.28	48.92	20.73	41.02
	PEGASUS$_{GPT}$	43.71	18.68	35.07	46.28	20.54	39.10
	BRIO$_{GPT}$	47.37	21.30	38.55	50.03	21.96	41.73
Human	GPT-4	24.95	5.64	18.57	36.80	10.90	32.05
	BART$_{GPT}$	26.39	6.61	19.10	40.05	14.86	35.08
	PEGASUS$_{GPT}$	28.00	7.94	20.77	40.50	16.18	35.76
	BRIO$_{GPT}$	26.81	7.01	19.81	40.39	15.19	35.20

Table 6: Evaluation of ROUGE Scores after Fine-Tuning with 20,000 GPT-4 Summaries. This table presents the calculated ROUGE scores, comparing various **Hypotheses** with **References**.

ing ROUGE scores for evaluation as shown in Table 6. This analysis contrasts human-generated summaries with those generated from GPT-4, noting lower ROUGE scores when comparing GPT-4 outputs to human references, highlighting differences in sentence structure and expression. Our results indicate variability in model performance, with GPT's BRIO model leading in ROUGE-1 and ROUGE-L scores on CNNDM, and GPT-based models surpassing human performance on XSum in these scores. Despite this, a significant performance gap exists between the best models and human summaries, particularly on XSum's ROUGE-2 scores. This result shows the strength of GPT-based models in abstract text generation, despite the challenges in closely mimicking human summarization.

5 Comparative Study

5.1 Training Efficiency

In Section 4.3, we detail the ROUGE score performance of various systems fine-tuned on a dataset of 20,000 GPT-4 generated references. The results show a discernible performance gap between the ROUGE scores achieved by our model and those reported in the original papers (Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2022), particularly concerning the XSum dataset. Therefore, we questioned whether fine-tuning the model on a larger dataset can yield further improvements in ROUGE performance. To check this, we created three sets of reference summaries from XSum articles using GPT-4, each varying in size, to serve as an enlarged training corpus. The specifics of the three datasets are detailed in Table 7.

First, we trained the model starting from the checkpoint fine-tuned on XSum, employing the same experimental setup as detailed in 3.4, results are reported in Table 8. On analysis, it becomes evident that augmenting the size of the dataset leads to an improvement in model performance, as measured by the ROUGE metric.

However, as we use the XSum checkpoint for its proven quality as a baseline, human reference remains crucial in our training process, leading to redundancy compared to other systems. To address this redundancy, we conducted additional experiments where, alongside using the XSum checkpoint, we initiated training with pre-trained weights for each model in this new configuration.

Variant	Train	Validation	Test
Small (20k)	20,000	1,100	11,334
Medium (50K)	50,000	2,750	
Large (100K)	100,000	5,500	

Table 7: Details of three dataset variations on XSum

System	Dataset	R-1	R-2	R-L
BART$_{GPT}$	Small	45.36	19.59	36.28
	Medium	47.44	21.47	38.34
	Large	**48.52**	**22.42**	**39.57**
PEGASUS$_{GPT}$	Small	43.71	18.68	35.07
	Medium	46.63	20.99	38.12
	Large	**47.62**	**22.13**	**39.32**
BRIO$_{GPT}$	Small	47.37	21.30	38.55
	Medium	48.82	23.28	40.66
	Large	**49.05**	**23.81**	**41.20**

Table 8: Evaluation of ROUGE scores after fine-tuning from the XSum checkpoint with various data sizes.

System	Dataset	R-1	R-2	R-L
BART$_{GPT}$	Small	46.28	20.37	37.26
	Medium	48.06	22.11	39.60
	Large	**48.84**	**23.13**	**40.68**
PEGASUS$_{GPT}$	Small	45.04	19.59	36.26
	Medium	47.21	21.76	38.80
	Large	**47.88**	**22.50**	**39.64**
BRIO$_{GPT}$	Small	47.64	21.68	38.93
	Medium	48.99	23.35	40.79
	Large	**49.33**	**24.08**	**41.44**

Table 9: Evaluation of ROUGE scores post fine-tuning from pre-trained weight with different data sizes.

Table 9 shows the result of fine-tuning from pre-trained weight. We observed that:

(1) The model performance can indeed be advanced by training with only LLM reference, **which proved that our dataset can substitute the original XSum dataset in the training procedure**.

(2) Compared to the model fine-tuned on the XSum checkpoint, the model that was fine-tuned from pre-trained weights demonstrated enhanced performance on identical data volumes. This improvement likely originates from variances between human reference and LLM reference (detailed in section 4.3), prompting the model to perceive previously trained targets as potential noise.

(3) **Our dataset reduces the performance gap across models like BART, PEGASUS, and BRIO**, indicating that summaries generated using LLM effectively counteract biases associated with the varied styles of human writers in the original dataset. Therefore, these LLM-generated summaries facilitate a smoother learning process for models, thereby diminishing the requirement for intricate training methodologies.

5.2 Novelty Analysis

In this section, we delve into the comparative analysis of novelty between the summaries authored by humans and those generated by GPT-4. Novelty is defined through the computation of novel n-grams, a method that serves to gauge the 'abstraction' of our models. The novelty metric is calculated[8] using the formula from Liu et al. (2022), i.e.,

$$Novelty(D, S^*) = \frac{\sum_{g \in G_{S^*}} \mathbb{1}(g \notin G_D)}{|G_{S^*}|} \tag{1}$$

where D and S^* are the source document and reference summary respectively, G_D and G_{S^*} are the sets of bigrams in D and S^*, $\mathbb{1}$ is the indicator function.

As presented in Table 10, models referencing GPT-4 exhibit better abstraction compared to those referencing human-generated summaries in the CNNDM dataset. Conversely, for the XSum dataset, models using human references are more "abstract" than those based on GPT-4 references. Despite these differences, as discussed in Section 4, summaries guided by GPT-4 are favored by human annotators across both the XSum and CNNDM datasets. This preference suggests that GPT-4, alongside our model, successfully balances the use of a diverse vocabulary for summary composition with effective information extraction from the original articles. Such a balance enhances summary relevance and aligns more closely with human preferences in summary generation.

	XSum		CNNDM	
System	**Unigram**	**Bigram**	**Unigram**	**Bigram**
Human	**.3399**	**.8342**	.1180	.4960
GPT-4	.2960	.8009	**.2375**	**.7074**
BART	**.2461**	**.7310**	.0118	.0922
BART$_{GPT}$	.1986	.6643	**.1287**	**.5389**
PAGASUS	**.2664**	**.7474**	**.1666**	.2919
PAGASUS$_{GPT}$	.1558	.5780	.0946	**.4616**
BRIO	**.2696**	**.7654**	.0258	.2261
BRIO$_{GPT}$	.2203	.7039	**.0962**	**.4890**

Table 10: Ratio of novel n-grams of various models on XSum and CNNDM. Novel n-grams are those that appear in the summaries but not in the source documents.

6 Conclusion

In this work, we propose a novel supervised learning framework using LLM-generated summaries as references. Our human evaluation compared systems guided by human-written and LLM-generated summaries. Results show LLMs can guide small models to produce summaries aligned with human preferences, opening new research directions in automatic summarization. We're releasing GXSum datasets in three sizes, containing XSum articles and LLM-generated summaries, which our experiments validate as potential replacements for the original XSum dataset. Our findings and dataset offer unique insights into LLM-enhanced automatic text summarization, encouraging further exploration of applying LLM knowledge to improve smaller, task-specific language models.

[8] The calculation is performed using ExplainaBoard (Liu et al., 2021). https://github.com/neulab/ExplainaBoard, and we had not employed PTBTokenizer prior to this calculation.

7 Limitations

Our work introduces a new dataset, GXsum, for which we employ summaries generated by GPT-4 as references. It is essential to note that in our experiments, summaries were generated using OpenAI's API, which, due to its rapid iteration capability, might result in variable outcomes that could limit the reproducibility of our experiments. Furthermore, constrained by the performance of GPT-4, the generated summaries may still possess a certain level of hallucination. Additionally, considering effectiveness, the dataset and generated summaries used in this experiment are confined to the news domain. Employing datasets from other domains might provide a more comprehensive analysis, which represents a potential future research direction for us. Lastly, the human evaluation experiments conducted aim to explore a wide range of human reading preferences. The outcomes may vary depending on the timing of the assessment and the platform used to employ evaluators; we merely state the observed facts.

References

Toufique Ahmed and Premkumar Devanbu. 2023. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, New York, NY, USA. Association for Computing Machinery.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 1–12.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. TrueTeacher: Learning factual consistency evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2053–2070, Singapore. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Junmo Kang, Wei Xu, and Alan Ritter. 2023. Distill or annotate? cost-efficient fine-tuning of compact models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11100–11119, Toronto, Canada. Association for Computational Linguistics.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better.

Wei Li, Xinyan Xiao, Yajuan Lyu, and Yuanzhuo Wang. 2018. Improving neural abstractive document summarization with explicit information selection modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1787–1796, Brussels, Belgium. Association for Computational Linguistics.

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

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021. ExplainaBoard: An explainable leaderboard for NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2023c. On learning to summarize with large language models as references.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016a. 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, Berlin, Germany. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016b. 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, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Karl Pearson. 1907. *On further methods of determining correlation*, volume 16. Dulau and Company.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM Transactions on Data Science*, 2(1):1–37.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022a. Salience allocation as guidance for abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8640–8665, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv e-prints*, pages arXiv–2204.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

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

A LLM Summary Generation

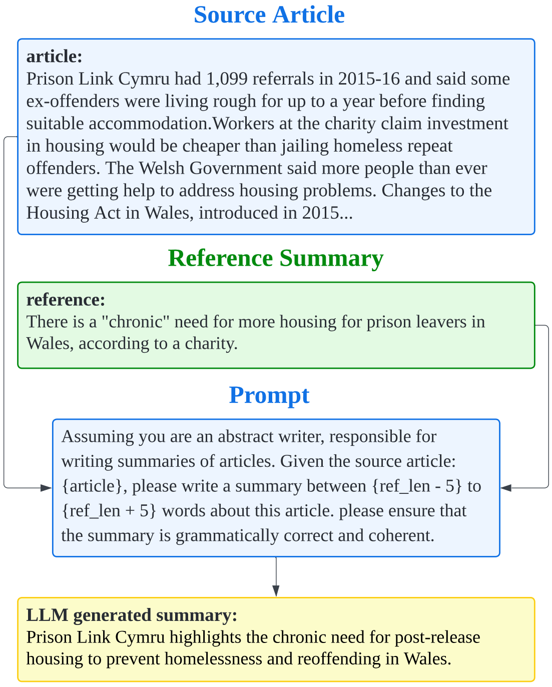
A.1 Prompt Template Example

Figure 3 illustrates the template for our prompt design. The {article} variable represents the source article from the original dataset, and the {len_lower} and {len_upper} variables represent the lower bound and upper bound length constraints that we will set.

> Assuming you are an abstract writer, responsible for writing summaries of articles. Given the source article: {article}, please write a summary between {len_lower} to {len_upper} words about this article. please ensure that the summary is grammatically correct and coherent.

Figure 3: Template for a ChatGPT API prompt.

A.2 Generation Process



Source Article

article:
Prison Link Cymru had 1,099 referrals in 2015-16 and said some ex-offenders were living rough for up to a year before finding suitable accommodation.Workers at the charity claim investment in housing would be cheaper than jailing homeless repeat offenders. The Welsh Government said more people than ever were getting help to address housing problems. Changes to the Housing Act in Wales, introduced in 2015...

Reference Summary

reference:
There is a "chronic" need for more housing for prison leavers in Wales, according to a charity.

Prompt

Assuming you are an abstract writer, responsible for writing summaries of articles. Given the source article: {article}, please write a summary between {ref_len - 5} to {ref_len + 5} words about this article. please ensure that the summary is grammatically correct and coherent.

LLM generated summary:
Prison Link Cymru highlights the chronic need for post-release housing to prevent homelessness and reoffending in Wales.

Figure 4: Illustration of LLM summary generation process

Figure 4 shows an example of our LLM summary generation process.

B Human Annotation Setting

B.1 Annotation Guideline

The definitions of various quality aspects we use in our annotation tasks are as follows:

- Relevance: Measures the importance of the summary content relative to the article, considering whether it has extracted the key points.

- Consistency: Considers whether the summary accurately includes all facts without fabricating false information.

- Fluency: Assesses whether each sentence in the summary is well-written and grammatically correct.

- Coherence: Considers whether the entire summary flows smoothly and reads naturally.

- Informativeness: Considers whether the summary clearly conveys the main message of the article, excluding unnecessary details.

B.2 Amazon Mechanical Turk Recruitment

To recruit qualified crowd annotators, stringent selection criteria were applied. These criteria stipulated that participants must have successfully completed more than 500 Human Intelligence Tasks (HITs), maintained an acceptance rate exceeding 95%, and resided within the United States. This rigorous selection process was implemented to guarantee that the annotators were native English speakers and had a demonstrable record of experience in effectively performing annotation tasks.

B.3 Upwork Recruitment

To ensure the rigor of expert analysis, we established specific criteria for selecting annotators, focusing on those with a requisite level of expertise. We engaged the Upwork platform to identify suitable candidates, stipulating prerequisites such as residency in English-speaking countries (specifically the USA, UK, Australia, or Canada), minimum educational attainment of a bachelor's degree, and prior experience in data annotation or linguistics-related roles. The ultimate selection of our expert candidates comprised individuals with backgrounds as writers, journalists, and professional text data annotators.

C Case Study

Article ID	36043765	AVG. Score
Human	A seal found tangled in nets on an Aberdeenshire beach has been returned to the sea.	1.0
GPT-4	Scottish SPCA rescues and frees a heavily netted grey seal at Cruden Bay, preventing potential lethal injuries.	3.0
BART$_{GPT}$	A large seal entangled in netting at Cruden Bay was rescued by the Scottish SPCA and safely released.	3.3
PEGASUS$_{GPT}$	Scottish SPCA rescued a grey seal from Cruden Bay with a large net tangled around its neck, releasing him unharmed.	3.0
BRIO$_{GPT}$	A grey seal entangled in heavy netting at Cruden Bay was rescued by the Scottish SPCA and released unharmed.	4.6

Table 11: Case study on XSum

Article ID	38537698	AVG. Score
Human	The reaction from BT's investors told us much about media regulator Ofcom's ruling on the fate of Openreach, the BT subsidiary that provides much of the UK's broadband infrastructure.	1.6
GPT-4	BT's shares surged after Ofcom ruled out a company break-up due to practical challenges, including a complex pension scheme and legal obstacles, potentially leading to future operational conflicts.	3.3
BART$_{GPT}$	Ofcom's Sharon White admits BT's potential break-up faces practical hurdles due to land deals, pension scheme complexities, and potential conflicts over Openreach's ownership and governance.	4.3
PEGASUS$_{GPT}$	BT shares rise 3% after Ofcom's decision not to break up the company, citing pension issues and land deals.	3
BRIO$_{GPT}$	Ofcom postpones BT's break-up due to pension scheme and land deals, prompting plans for separation and potential conflicts of interest.	2.6

Table 12: Case study on XSum

Article ID	ee17dfb574feca82ccac5689595e47483bd23f12	AVG. Score
Human	London's Metropolitan Police say the man was arrested at Luton airport after landing on a flight from Istanbul. He's been charged with terror offenses allegedly committed since the start of November.	1.0
GPT-4	British man, Yahya Rashid, 19, faces terror charges upon his UK return from Turkey. Arrested at Luton airport, he's accused of preparing and aiding terrorist acts from November to March. Rashid will appear in court in Westminster.	2.6
BART$_{GPT}$	Yahya Rashid, a 19-year-old from northwest London, was charged with terror offenses after his arrest at Luton airport on his return from Turkey. He faces charges of terrorism preparation and aiding acts of terrorism between November 1 and March 31, with a court appearance set for Wednesday.	5.0
PEGASUS$_{GPT}$	Yahya Rashid, a 19-year-old from London, was charged with terrorism offenses at Luton Airport after returning from Turkey. He faces charges of preparing acts of terrorism and assisting others to commit terrorism.	2.6
BRIO$_{GPT}$	19-year-old Yahya Rashid, a UK man, was charged with terror offenses after his arrest at London's Luton airport after his return from Turkey. He faces charges for planning and aiding acts of terrorism between November 1 and March 31, with his court appearance set for Wednesday.	3.6

Table 13: Case Study on CNNDM.