Towards Enhancing Coherence in Extractive Summarization: Dataset and Experiments with LLMs

Anonymous ACL submission

Abstract

Extractive summarization plays a pivotal role in natural language processing due to its wide-range applications in summarizing diverse content efficiently, while also being faithful to the original content. Despite significant advancement achieved in extractive summarization by Large Language Models (LLMs), these summaries frequently exhibit incoherence. An important aspect of the coherent summary is its readability for intended users. Although there have been many datasets and benchmarks proposed for creating coherent extractive summaries, none of them currently incorporate user intent to improve coherence in extractive summarization. Motivated by this, we propose a systematically created human-annotated dataset consisting of coherent summaries for five publicly available datasets and natural language user feedback, offering valuable insights into how to improve coherence in extractive summaries. We utilize this dataset for aligning LLMs through supervised fine-tuning with natural language human feedback to enhance the coherence of their generated summaries. Preliminary experiments with Falcon-40B and Llama-2-13B show significant performance improvements ($\sim$ 10% Rouge-L) in terms of producing coherent summaries. We further utilize human feedback to benchmark results over instruction-tuned models such as FLAN-T5 which resulted in several interesting findings[1].

1 Introduction

With the increasing amount of information, the significance of automatic summarization has grown exponentially. Summarization techniques can be broadly classified into two categories: (i) Extractive, and (ii) Abstractive. The abstractive methods (Nallapati et al., 2016; Gupta, 2019) often focus on the semantic meaning of the text, giving a summary by creating a new set of sentences. However,

[1]Data and source code are available at <anonymous link>

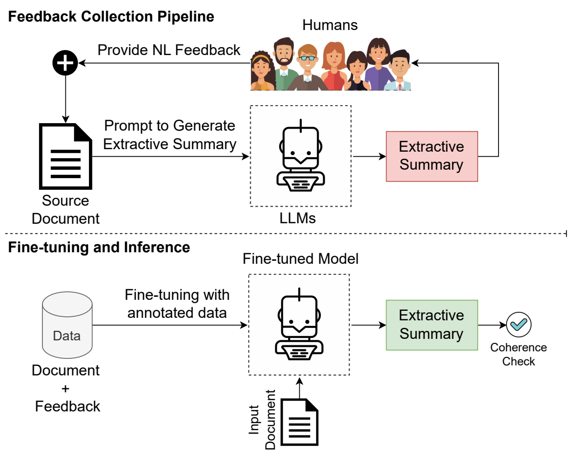


Figure 1: Schematic representation of our natural language feedback collection pipeline and aligning LLMs with provided human feedback.

these methods often struggle with generating ungrammatical or even nonfactual contents (Kryscinski et al., 2020; Zhang et al., 2022). In contrast, extractive methods focus on selecting meaningful phrases/sentences from the given text, giving a summary that is faithful to the original content, hence it has a range of real-world applications (Zhang et al., 2023a). For instance, tasks such as video shortening, and legal document summarization require precision and adherence to specific details from original text, and extractive methods are more suitable for these tasks. However extractive summarization often generates summaries that lack coherence, and coherence is a crucial attribute of text summarization since it holds a significant connection to user experience. Thus, our work aims to improve coherence in extractive summarization.

With the advent of LLMs such as GPT-4, Llama-2 (Touvron et al., 2023), and Falcon (Penedo et al., 2023), there is a significant advancement in generating extractive summaries (Zhang et al., 2023a; Stiennon et al., 2020). For extractive summarization, coherence is often measured through the interconnection among sentences and ease of readability for users. Past attempts have been made to improve and quantify coherence in extractive summarization (Nallapati et al., 2016; Wu and Hu, 2018; Jie et al., 2023a)[2], however, these attempts do not consider user-specific intent (i.e., ease of readability while preserving important information). Thus, we approach the concept of coherence through the lens of user-specific intent (Figure 1). To this end, we propose a comprehensive dataset with a systematic collection of natural language feedback to improve coherence in model-generated summaries, and human-annotated extractive coherent summaries. To the best of the authors' knowledge, this dataset represents the initial effort to align the coherence in a summary with user intent.

To develop the proposed dataset, we hire expert annotators to accurately annotate data for our task. For the annotation, the objective is two-fold: (1) to create a coherent summary by extracting important sentences from a source document that effectively captures the key aspects of the document, and (2) to provide feedback (i.e, natural language explanations) on the steps to go from the model summary to the gold coherent summary. We annotate this data across five categories: News, Debate, TV Show, Meeting, and Dialogue. Our annotation process consists of three phases (detailed discussion in §2). Each data instance collected in our dataset consists of <*Source text, Initial model summary, Feedback, Gold coherent summary, Scores*> elements.

We utilize the proposed dataset for aligning widely used open-source LLMs to generate more coherent extractive summaries via supervised fine-tuning: (i) two decoder-only models, i.e., Falcon-40B and Llama-2-13B, and (ii) three encoder+decoder models, i.e., FLAN-T5, Tk-Instruct, and T5. We develop a baseline and propose two different supervised fine-tuning strategies with human feedback (details are presented in §3). We measure the performance in terms of Rouge-L. Rouge-L assesses the syntactic and semantic similarity between the generated and the gold coherent summary, indicating their proximity. We also provide human judgments in terms of the coherence of generated summaries by baseline and proposed approach. Experimental results reveal that the proposed models show absolute improvement of $\sim 10\%$ Rouge-L over baselines. Furthermore, human evaluation shows a preference for extractive summaries from our approach, often rating them as more coherent. This indicates that aligning the model with user feedback improves coherence. Furthermore, a thorough analysis of the results reveals several interesting findings. We hope that our findings facilitate future research for improving coherence in extractive summarization.

2 Data Collection

Our annotation process consists of three phases. First, we randomly select a source text for annotation across five different categories from publicly available datasets. Second, we prompt a large language model to create coherent summaries for selected source text. Finally, we hire expert annotators to review generated summaries and provide natural language feedback/explanations to improve coherence in generated summaries.

2.1 Source Datasets

Our comprehensive annotated dataset consists of five different categories: News, Debate, TV Show, Meeting, and Dialogue. We carefully curated data for each category by randomly selecting 200 instances from publicly available datasets. In particular, we exclusively utilize the input/source text for annotation purposes from all of these datasets. We leverage CNN/DM dataset (Nallapati et al., 2016) for news, DebateSum (Roush and Balaji, 2020) for Debate, TVQA (Lei et al., 2018) for TV Show, MeetingBank (Hu et al., 2023) for Meeting, and DialogueSum (Chen et al., 2021) for Dialogue category. Further details are presented in App. C.

2.2 Coherent Summary Generation

The objective is to generate an extractive summary, where the model is prompted to select the most suitable sentences from the document for coherent summarization. Thus, we formulate an extractive summarization task as selecting sentences from a given document to produce coherent summaries. Let us consider document $\mathcal{D}$. We first divide $\mathcal{D}$ at the sentence level and create set $\mathcal{D}_s = \{s_1, s_2, ..., s_n\}$, where s_i denotes the i^{th} sentence from $\mathcal{D}$. To create numbered sentences from the document, we use the NLTK library[3]. Now, we prompt (p) the Falcon-40B-Instruct model (denoted as $\mathcal{M}$) to produce a coherent summary from the source text provided as $\mathcal{D}_s$. To accomplish this, we employ a

[2]Detailed related work is presented in App. B

[3]https://www.nltk.org/api/nltk.tokenize.html

1-shot prompting approach (prompt is presented in the App. A). Formally, we present our task as $\mathcal{M}(p, \mathcal{D}_s) = C_s$, indicates that the task for $\mathcal{M}$ is to produce coherent summary (denoted as C_s) by selecting sentences from $\mathcal{D}_s$ given p.

2.3 Annotation Process

We use the Upwork platform to hire expert annotators to annotate our dataset. We initiated a pilot project involving 25 annotators having a strong background and fluency in the English language. Evaluating their performance during the pilot phase, we subsequently hired 10 proficient annotation experts to carry out the final annotations. Annotators are provided with task instructions, source text, and model summary (generated in §2.2). They are expected to produce a coherent summary based on the provided source text by selecting sentences/phrases from the document and provide feedback on the steps to go from the model summary to the gold coherent summary (annotated by them). Each source text is annotated by 3 different annotators. Along with that, they need to rate the model summary based on three criteria (i.e., Relevance, Coherence, and Consistency) on a Likert scale of 1-5, motivated by Fabbri et al. (2021). A annotated data instances consist of five elements as illustrated in Figure 2. A detailed example and further annotator details are presented in App. D.

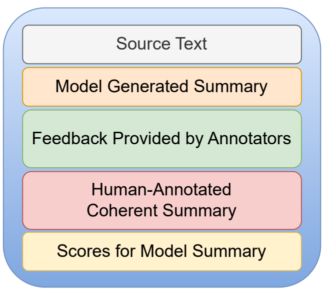


Figure 2: Illustration of annotated instance

Source text is the document provided to annotators which falls under one of five categories.

Model-generated summary The summary generated in §2.2 is provided to annotators.

Coherent Summary is generated by annotators from the given source document.

Feedback is a natural language explanation provided by annotators to improve coherence in the model summary and achieve a coherent summary generated by them.

Scores Annotators score the model-generated summary to measure the three different aspects: (i) Relevance: measure the selection of important content (key points) from the source, and the summary should include only important information from the source document; (ii) Coherence: measure the collective quality of all sentences, and the summary should be well-structured and well-organized; and (iii) Consistency: measure the factual consistency of the summary that contains only statements that are entailed by the source document.

2.4 Quantitative Analysis

Annotators have annotated a total of 1000 unique samples and each sample is annotated by three different annotators with the inter-annotator agreement of 0.659 (details in App. D.2). For each document category, 200 samples are annotated. After all annotations, the average scores for model summary are: (1) Relevance: 3.81, (2) Coherence: 3.46, (3) Consistency: 4.09. Here, coherence is low for the model-generated summary which suggests that improving coherence is essential task.

3 Experiments and Results

3.1 Experimental Setup

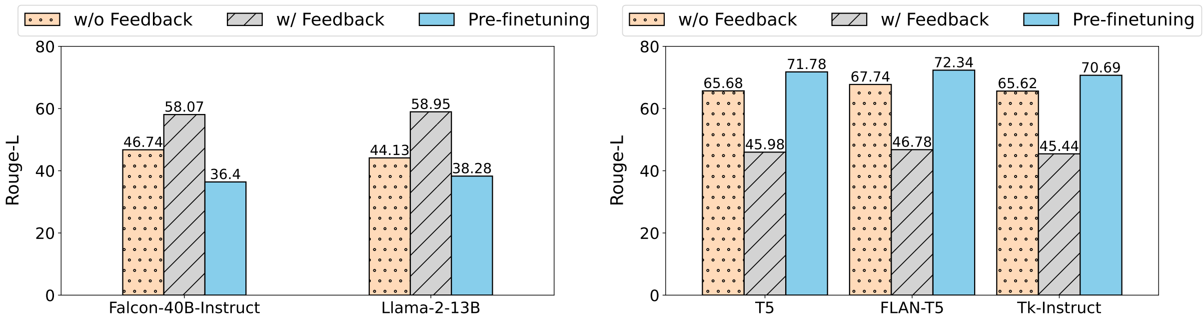
Models We perform experiments with five different models with two architecture families: (i) two Decoder (Dec.) only open-source LLMs (Falcon-40B, and Llama-2-7B), and (ii) three Encoder (Enc.) + Decoder (Dec.) models (T5-large, and two instruction-tuned models, FLAN-T5-large and Tk-Instruct-large). In experiments, Dec. only models are fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2021), and Enc.+Dec. models are fine-tuned using full-parametric training. We employ three different strategies to fine-tune these models.

Baseline fine-tuning model on <*Source text*> as input and <*Coherent Summary*> as output.

w/ Feedback fine-tuning model on <*Source text, Initial model summary, Feedback*> as input and <*Coherent Summary*> as output.

Pre-finetuning First, we fine-tune the models on <*Source text*> as input and <*feedback*> as the output. Subsequently, we execute supervised fine-tuning by employing <*Source text*> as the input and <*Coherent Summary*> as the output on the pre-finetuned model.

Our approaches reflect an effort to refine the models' coherence by leveraging feedback and



(a) (b)

Figure 3: Performance of (a) Dec. only model, and (b) Enc. + Dec. Model on our proposed dataset.

user-driven insights during the fine-tuning. We fine-tune the model to generate sentences as a summary (format of the coherent summary is shown in Table 2) which ensures the extractive nature of generated summaries. The dataset is randomly divided into train (80%), and test (20%) sets. For comparability, we use the same hyperparameter settings for all runs: trained for 3 epochs, with a batch size of 16 and an initial learning rate of 5e-5. All experiments were conducted on A100 NVIDIA GPUs.

Metric We use Rouge-L (Lin, 2004) to evaluate model performance by measuring the similarity between the generated summary and the gold standard coherent summary. Our assessment is based on how closely the model summary resembles this gold standard, indicating coherence similarity. To supplement this objective measure, we also perform human evaluations of the generated summaries.

3.2 Results and Analysis

Here, we compare the baselines and proposed methods despite different fine-tuning approach since the inference is consistent: <*Source text*> is input, and <*Coherent Summary*> is output. Models do not have access to feedback during inference.

Effect of Feedback on Dec. only models Figure 3a shows the Rouge-L scores for Falcon-40B-Instruct and Llama-2-13B, comparing baseline and proposed methods. The proposed methods, involving fine-tuning with user feedback, clearly outperform the baselines: Falcon improves by 11.33%, and Llama by 14.82%. However, both models' performance drops significantly during pre-finetuning with feedback data. This pre-finetuning aims to integrate feedback knowledge into the model's parameters. When fine-tuning with LoRA, updating only the adaptation layer, performance decreases during pre-finetuning. However, the efficacy of pre-finetuning becomes evident with full-parametric training, as shown in Figure 3b.

Effect of Feedback on Enc. + Dec. models Figure 3b represents the Rouge-L scores for FLAN-T5, Tk-Instruct, and T5, comparing both baseline and proposed methods. From results, it becomes evident that directly fine-tuning with user feedback doesn't enhance the performance of these models as shown with Dec. only models. Conversely, adopting a pre-finetuning enhances the performance of these models significantly (further discussion in App. E). Figure 3b shows that pre-finetuning leads to improved performance, with the T5, FLAN-T5, and Tk-Instruct models surpassing baseline by 6.1%, 4.6%, and 5.07%, respectively.

Human Evaluation We aim to examine the correlation between human judgments and Rouge-L. To this end, we conduct a case study involving human evaluation presented in App. E.

4 Conclusions

This paper introduced a comprehensive dataset designed to improve coherence in extractive summarization while integrating natural language feedback from human users across five different categories. Utilizing this dataset, we conducted evaluations using various LLMs, and initial experimental outcomes demonstrate an enhancement in model performance, with $\sim 10\%$ improvement in coherence achieved through fine-tuning with human feedback. Moreover, our analysis highlights the potential for performance advancements in instruction-tuned models through pre-finetuning based on user feedback. We believe that both the dataset and the findings derived from this work will serve as valuable tools for future research in this direction.

Limitations

Though we evaluated our approach on a widely-used range of LLMs including Falcon-40B and LLaMa-2-7B, this study can also be extended to other LLMs. To improve the utilization of human feedback collected in our dataset, development of advanced methods such as iterative feedback loops and dynamic feedback during both training and inference stages can be interesting future research direction. Since manual annotation of feedback is time-consuming and laborious, exploration of automated methods for feedback generation using smaller-scale supervised learning or LLMs is necessary. Additionally, we hope to expand our analysis to include the most recent LLMs such as GPT-4 and ChatGPT on our proposed dataset. We also note that this research is limited to the English language and can be extended to multilingual scenarios for improving coherence in extractive summarization.

Ethics Statement

We have used AI assistants (Grammarly and ChatGPT) to address the grammatical errors and rephrase the sentences.

References

Mohamad Abdolahi and Morteza Zahedi. 2019. Textual coherence improvement of extractive document summarization using greedy approach and word vectors. *International Journal of Modern Education and Computer Science*.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1173, Atlanta, Georgia. 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.

Som Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A benchmark dataset for meeting summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Litton J Kurisinkel, Pruthwik Mishra, Vigneshwaran Muralidaran, Vasudeva Varma, and Dipti Misra Sharma. 2016. Non-decreasing sub-modular function for comprehensible summarization. In *Proceedings of the NAACL Student Research Workshop*, pages 94–101, San Diego, California. Association for Computational Linguistics.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2023a. Enhancing coherence of extractive summarization with multitask learning. *arXiv preprint arXiv:2305.12851*.

Renlong Jie, Xiaojun Meng, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Enhancing coherence of extractive summarization with multitask learning. *ArXiv*, abs/2305.12851.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Litton J Kurisinkel and Vasudeva Varma. 2015. Readable and coherent multidocument summarization.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 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, Barcelona, Spain. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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, Berlin, Germany. Association for Computational Linguistics.

Daraksha Parveen and Michael Strube. 2015. Integrating importance, non-redundancy and coherence in graph-based extractive summarization. In *International Joint Conference on Artificial Intelligence*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Allen Roush and Arvind Balaji. 2020. DebateSum: A large-scale argument mining and summarization dataset. In *Proceedings of the 7th Workshop on Argument Mining*, pages 1–7, Online. 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.

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

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535, Seattle, United States. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023b. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

A Prompt

In this section, we provide an example of a 1-shot prompt used in §2.2. The prompt consists of the task definition, one example, and an input instance.

Task

You are an extractive summarizer. You are presented with a document. The document is a collection of sentences and each sentence is numbered with sentence ids. Understand the given document and create a meaningful summary by picking sentences from the document. Please list the sentence IDs as output so that sentences corresponding to the generated IDs summarize the document coherently.

Example

Learn from the below example:
Document:
1. Olympic gold medallist Jessica Ennis-Hill has confirmed she will return to competition in London this July following her break from athletics to become a mother.
2. Ennis-Hill provided one of London 2012's most captivating storylines by surging to heptathlon gold, and the Sheffield-born star will return to the Olympic Stadium three years on to compete in the Sainsbury's Anniversary Games.
3. The 29-year-old has not competed since the same event in 2013 and gave birth to her son, Reggie, last summer.
.
.
.
13. Ennis-Hill will take part in the two-day meeting on July 24 and 25, with the Sainsbury's IPC Athletics Grand Prix Final taking place on July 26.
14. Ennis-Hill added: 'The 2012 Olympics were an incredible experience for me and it will be very special to step out on that track again.
15. It will be amazing to compete in front of all our British fans who I am sure will have their own memories of the London Games too.

Summary: <s> [2, 5, 6, 11, 12, 15]

Input

Document: [source text]
Please Create a concise summary using as few sentences as possible.

Summary: <s>

The example given in this prompt is annotated by the authors where we reviewed the document and chose specific sentence IDs to create a coherent summary.

B Related Work

There are some past attempts that have been made to improve coherence in extractive summarization. Christensen et al. (2013) proposed a G-FLOW, a joint model for selection and ordering sentences that balances coherence for multi-document extractive summarization. After that, Parveen and Strube (2015) proposed a graph-based method for extractive single-document summarization that considers importance, non-redundancy, and local coherence simultaneously. In addition, Kurisinkel and Varma (2015) introduced A multi-document summarization method that ensures content coverage, sentence ordering, topical coherence, topical order, and inter-sentence structural relationships using a Local Coherent Unit (LCU). Following this, J Kurisinkel et al. (2016) proposed scoring-based function to identify the discourse structure which provides the context for the creation of a sentence for generating comprehensible summaries. Furthermore, Wu and Hu (2018) utilized reinforcement learning to extract a coherent summary, and Abdolahi and Zahedi (2019) enhanced coherence in extractive document summarization through a greedy approach and word vectors. In addition, Jie et al. (2023b) introduced two strategies, including pre-trained converting models (model-based) and converting matrices (MAT-based) that merge sentence representations to improve coherence. With the emergence of LLMs, Zhang et al. (2023b) attempted to analyze the performance of GPT-3 with different prompting for generating coherent summaries. Differing from these existing efforts, we approach the concept of coherence within summaries through the lens of user-specific intent.

C Datasets

In this section, we discuss more details about publicly available datasets used for developing our proposed benchmark.

CNN/DM The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail (Nallapati et al., 2016). We utilize randomly selected 200 news articles from this dataset for our annotations.

DebateSum DebateSum is constructed from evidence related to annual policy debate resolutions (Roush and Balaji, 2020), each averaging around 560 words. As DebateSum spans seven years of content, it encompasses seven distinct resolutions. For our annotations, we randomly selected 200 resolution plans from this dataset.

TVQA TVQA is a large-scale video QA dataset based on 6 popular TV shows (Friends, The Big Bang Theory, How I Met Your Mother, House M.D., Grey's Anatomy, and Castle) (Lei et al., 2018). From this dataset, we utilize subtitles-based dialogues as source text for our annotation.

MeetingBank MeetingBank is a benchmark dataset created by the city councils of 6 major U.S. cities to supplement existing datasets. It contains 1,366 meetings with over 3,579 hours of video, as well as transcripts, PDF documents of meeting minutes, agenda, and other metadata (Hu et al., 2023). From this dataset, we utilize transcripts as source text for our annotation.

DialogueSum DialogSum is a large-scale dialogue summarization dataset, consisting of 13,460 dialogues with corresponding manually labeled summaries and topics (Chen et al., 2021). We utilize randomly selected 200 dialogues from this dataset for our annotations.

D Example of Annotated Instance

In this section, we provide an example of an annotated data instance from the News category in Table 2. This instance provides an illustrative example of how the whole dataset is collected. We also conduct analysis of the collected data focusing on how improving coherence affects the length of summaries, offering insights into the impact on the length of summaries. We observed that the average lengths of the original documents, model-generated summaries, and coherently annotated summaries are 24.89, 17.99, and 11.95 sentences, respectively. These findings suggest that annotators often removed sentences to enhance the coherence of the summaries during the annotation process.

D.1 Annotator Details

Our annotators consist of contractors hired through Upwork. Annotation of each data instance paid \$3 and could be completed within 20 minutes, compensating an annotator with an average pay of \$15/hour. The final annotation process took around time of $\sim$ 15 days and cost of $\sim$ $\$10k$. Overall, we collected a total of 1000 unique samples, and the dataset was randomly partitioned into training (80%), and test (20%) sets. We also provide the final 10 annotators' demographic data in terms of their nationality in Table 1.

Nationality	# of Annotators
India	3
Philippines	3
Venezuela	1
Pakistan	1
Macedonia	1
Kenya	1

Table 1: Demographic details of annotators

D.2 Calculation of Inter-annotator Agreement

To calculate the inter-annotator agreement using ROUGE for three annotators, we focused on the ROUGE-L metric, which measures the longest common subsequence between summaries. Since the extractive summaries they have annotated are selections of sentences from the article, it makes sense to use ROUGE-L to capture the structural similarity of their selections. For each document, we computed the ROUGE-L score for every possible pair of annotators, capturing the consistency of their sentence selections. By averaging these pairwise ROUGE-L scores across all documents, we obtained an overall agreement score that reflects how closely the annotators' summaries align in terms of content and structure. This approach provides a quantitative measure of agreement that highlights the consistency among annotators in annotating the extractive summaries.

E Extended Discussion on Analysis

Performance of encoder-decoder *vs.* decoder-only models The observed differences in the impact of feedback on encoder-decoder models *vs.* decoder-only models can be attributed to pre-training methodologies for both types of models. Encoder-Decoder models (e.g., T5, FLAN-T5) are pre-trained using a sequence-to-sequence framework, where the encoder processes the input text and the decoder generates the output text (Raffel et al., 2020). Decoder-only models (e.g., Falcon-40B, Llama-2) are pre-trained using a left-to-right autoregressive approach, predicting the next token based on the preceding tokens (Radford et al., 2019). When models are fine-tuned on <Source

text, Initial model summary, Feedback>, decoder-only models benefit more compared to encoder-decoder models because the feedback helps them align their sequential generation process more closely with human corrections. The pre-finetuning approach involves an intermediate step where models are first fine-tuned on <Source text> as input and <feedback> as the output. For encoder-decoder models, this step helps integrate feedback more effectively into their bidirectional context understanding, leading to significant improvements. For decoder-only models, this approach does not always yield better results as they benefit more directly from feedback fine-tuning. In summary, the differential impact of feedback on encoder-decoder and decoder-only models can be attributed to their respective pre-training objectives.

Human Evaluation We asked three independent human evaluators (graduate student volunteers) to assess the summaries (50 randomly selected from the test set). Each evaluator was asked to choose their preferred summary from three options: (1) the model summary (provided during annotations), (2) Llama-2 (w/o feedback), and (3) Llama-2 (w/ feedback). Additionally, they were asked to rate each summary's coherence on a Likert scale ranging from 1 (incoherent) to 5 (perfectly coherent). We calculate the inter-annotator agreement based on their choice of preferred summary. Since coherence is very subjective to annotators, we found 0.513 inter-annotator agreement (measured with raw/observed agreement) between three different annotators.

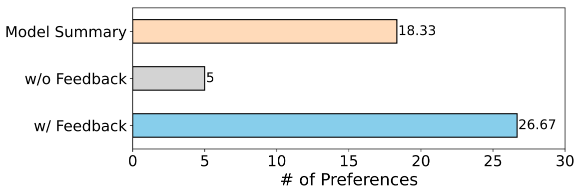


Figure 4: Average number of preferences across three evaluators.

Figure 4 shows the results for an average number of preferences across three evaluators, and the average coherence score is 3.45, 2.29, and 3.53 for model summary, Llama-2 (w/o feedback), and Llama-2 (w/ feedback), respectively. The results revealed that, on average, the evaluators favored the summary from Llama-2 (w/ feedback), which also received the highest average coherence score. These findings are consistent with and further corroborated by the results presented in Figure 3a. This further supports the findings presented in the paper using Rouge-L.

Document: If anyone won this debate it was the women. Their less choreographed style of body language gave the impression we were listening to real messages from real people rather than watching spin doctors' puppets performing. Overall I'm sure Miliband's coaching team will be patting themselves on the back and ....
Model Summary: Their less choreographed style of body language gave the impression we were listening to real messages from real people rather than watching spin doctors' puppets performing. Nicola Sturgeon (pictured) is a smiling assassin, ...
Coherent Summary: Sent. 1: If anyone won this debate it was the women. Sent. 2: Their less choreographed style of body language gave the impression we were listening to real messages from real people rather than watching spin doctors' puppets performing. Sent. 3: In his après-Paxman mode, David Cameron (pictured) was looking serious and oozing leadership charisma . .....
Feedback: Sent. 1: If anyone won this debate it was the women. Feedback 1: Add this sentence to give an idea what the summary is all about. . . Sent. 6: Clegg is a good speaker but his performance was vintage, ie a complete re-run of his 2010 routine. Feedback 6: Add this sentence in the model summary to provide information about the speaker. . . Sent. 9: He took enough pops at Cameron and waved his arm enough in that direction to signal an official end to the relationship that began in the Rose Garden but he looked more congruent agreeing with Cameron or fielding criticism as a double act than he did turning on him, which looked rather panto. Feedback 9: Add this sentence in the model summary as a supporting detail to the previous sentence.
Scores: Relevance: 4 Coherence: 3 Consistency: 5

Table 2: Illustrative example of annotated instance. Certain text is redacted due to space constraints.