Disordered-DABS: A Benchmark for Dynamic Aspect-Based Summarization in Disordered Texts

Xiaobo Guo and **Soroush Vosoughi**
Department of Computer Science
Dartmouth College
Hanover, New Hampshire
{xiaobo.guo.gr, soroush.vosoughi}@dartmouth.edu

Abstract

Aspect-based summarization has seen significant advancements, especially in structured text. Yet, summarizing disordered, large-scale texts, like those found in social media and customer feedback, remains a significant challenge. Current research largely targets predefined aspects within structured texts, neglecting the complexities of dynamic and disordered environments. Addressing this gap, we introduce Disordered-DABS, a novel benchmark for dynamic aspect-based summarization tailored to unstructured text. Developed by adapting existing datasets for cost-efficiency and scalability, our comprehensive experiments and detailed human evaluations reveal that Disordered-DABS poses unique challenges to contemporary summarization models, including state-of-the-art language models such as GPT-3.5.

1 Introduction

The exponential growth of digital content has significantly increased the importance of automated text summarization methods. These methods are crucial for distilling salient content from large volumes of text and efficiently addressing diverse information needs. Query-focused summarization (QFS) (Wang et al., 2022; Zhong et al., 2021) and aspect-based summarization (ABS) (Hayashi et al., 2021; Ahuja et al., 2022) have emerged as prominent approaches for creating focused summaries based on specific aspects or queries. This advancement has led to the development of several specialized datasets and benchmarks.

While QFS and ABS are adept at generating summaries targeted at specific aspects, they typically depend on predefined aspects for querying or model fine-tuning. Dynamic Aspect-Based Summarization (DABS) was introduced to overcome this limitation. DABS identifies aspects dynamically from the content of source articles, transcending the restrictions of fixed aspect definitions.

Figure 1: An example from Reddit showcasing a multi-user discussion. Each participant's input is color-coded to distinguish the varied aspects they discuss, emphasizing the range of perspectives and topics within the conversation. This visual differentiation serves to highlight the diversity inherent in such online discussions.

Despite the innovations in DABS, a significant challenge remains, particularly in real-world applications: the blending of sentences from varied sources within a text. This complication is frequently observed in scenarios like summarizing community opinions, analyzing user feedback, or collating medical records from multiple sources. Such situations usually involve quick changes in topics or aspects, leading to a notable inconsistency in coherence among texts from different origins. Figure 1 illustrates a discussion from Reddit involving multiple users. The dialogue initially focuses on the appeal of New York City, transitions to its drawbacks for long-term residence, and eventually shifts to a conversation about Sao Paulo. This example underscores the dynamic nature of topics in online discussions, highlighting the complexity of summarizing such disordered texts [1]. This disorder not only makes the summarization process more complex but also requires strategies adept at seamlessly merging disparate aspectual data, adding another layer of complexity to the task.

[1]The full discussion can be seen on Reddit

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 416–431
November 12-16, 2024

Our paper addresses these challenges by presenting a novel and efficient method to modify existing summarization datasets, making them suitable for dynamic aspect-based summarization in disordered texts. We use this method to introduce the Disordered-DABS dataset[2], specifically curated for this demanding summarization task. Comprising 255,286 aspect-oriented summaries, this dataset is segmented into training, validation, and testing partitions. Its robustness is validated through comprehensive human evaluations. Our examination of various baseline models highlights the complexity of this task, revealing that even advanced large language models like GPT-3.5 face significant challenges in dealing with the nuances of Disordered-DABS. These observations underscore the intricacy inherent in this form of summarization, highlighting the necessity for more specialized datasets tailored to its unique requirements.

2 Related Work

QFS has long been a fundamental task in the field of text summarization, focusing on generating summaries tailored to specific queries (Dang, 2005). It targets extracting precise information relevant to a wide range of queries, reflecting the diverse informational needs in various contexts. Key datasets in QFS, such as those developed by Xu and Lapata (2021), have been instrumental in addressing these needs, employing queries like "steroid use among female athletes" to explore detailed aspects such as "trends", "side effects", and "consequences".

ABS emerged as an extension, initially focusing on customer review analysis (Hu and Liu, 2004; Lu et al., 2009). ABS expands the scope of summarization to a broader range of domains, including news articles and encyclopedia content, as evidenced by works in ABS tasks (Frermann and Klementiev, 2019; Ahuja et al., 2022) and dataset creation (Hayashi et al., 2021). This shift from query-specific summaries in QFS to aspect-oriented summaries in ABS demonstrates the evolution of summarization techniques in response to varying content and context requirements.

Further advancing this evolution is DABS, which diverges from traditional ABS by not limiting aspects to predefined categories. Innovations in DABS, such as AnyAspect (Tan et al., 2020) and ENTSUMv2 (Mehra et al., 2023), have focused on utilizing entities as aspects, with AnyAspect summarizing sentences around named entities and ENTSUMv2 employing human-annotated, entity-specific summaries. Other datasets like OASUM (Yang et al., 2023b) and OPENASP (Amar et al., 2023) explore conceptual labels or sub-topics as aspects, with OASUM leveraging Wikipedia subtitles and OPENASP adopting a multi-document DABS settings. A comparative overview of these datasets is presented in Table 1.

However, a notable limitation of existing datasets in both ABS and DABS is their primary focus on well-organized and coherent texts, such as a news article. This focus is less relevant to platforms such as social media, where content often features multiple authors and rapid shifts in context. The inherent coherence in these texts could potentially overstate the performance of ABS and DABS models, as they may depend on sentence transitions to discern changes in topics or aspects. This coherence within texts of the available datasets might inadvertently inflate the performance of ABS and DABS models, as they might rely on transitions in contiguous sentences to infer changes in topics or aspects, whereas, in a dataset from multiple sources, this transition may detect the change of author and not aspect. Therefore, we propose the Disordered-DABS in which the aspects are dynamic and the texts are disordered.

3 Dynamic Aspect-Based Summarization

Despite substantial efforts in QFS and ABS, the definition of "aspect" remains ambiguous. It ranges from different topics (Frermann and Klementiev, 2019; Maddela et al., 2022; Tan et al., 2020) to specific elements within a topic (Hayashi et al., 2021; Ahuja et al., 2022; Meng et al., 2021). Some studies use concise phrases for aspects (Amar et al., 2023; Angelidis et al., 2021), while others allow longer sentences for complex information needs (Zhong et al., 2022, 2021). Given this variability in aspect definition, our Disordered-DABS incorporates two distinct sub-datasets: one emphasizing aspect diversity across varied topics and another focusing on different aspects of a single event, providing a comprehensive benchmark for these divergent aspect interpretations.

Building on previous DABS research (Tan et al., 2020; Yang et al., 2023b), our task is defined as generating multiple aspect-based summaries from a document D containing n topics or aspects. Unlike traditional baselines that pair a document with

[2]The data is available on Github

Dataset	Domain	Disordered	Collection	# Instances
AnyAspect (Tan et al., 2020)	News	✗	automatic	312,085
OASUM (Yang et al., 2023b)	Wikipedia	✗	automatic	3,747,569
OpenASP(Amar et al., 2023)	News	✗	manual	1,310
ENTSUMV2(Maddela et al., 2022)	News	✗	automatic	2,788
Disordered-DABS (Ours)	News & Wikipedia	✓	automatic	255,286

Table 1: Prominent datasets for dynamic aspect-based summarization. "# Instances" is the number of data instances in the dataset. "Disordered" means the sentences are shuffled in the source article

a specific aspect for aspect-based summary generation (Amar et al., 2023), our task requires the model to autonomously generate multiple summaries without predefined aspects. Additionally, to challenge the model's ability to handle non-coherent texts, we shuffle the sentences within D, following Frermann and Klementiev (2019), disrupting the natural flow and requiring the model to cluster information from the entire document.

3.1 Automatic Evaluation Metrics

Automatic evaluation metrics are vital for the efficient development and evaluation of models for the ABS and DABS tasks. Our evaluation focuses on two primary factors: the capability to identify aspect-based information and the overall quality of the generated summaries. We measure the aspect identification accuracy using the absolute difference in aspects between the reference and generated summaries (#AbsAspDiff). For assessing summary quality, we employ the standard summary quality metrics Rouge-1/2/L (Lin, 2004) and BERTScore (Zhang et al., 2019).

To compare reference and generated summaries effectively, we pad the summaries to equalize the number of aspects. This approach penalizes models that either miss aspects or generate excessive ones. The optimal pairing of reference and generated summaries ($\langle \hat{S}, S \rangle$) is determined by maximizing the mean performance across all metrics, as defined by:

$$\langle \hat{S}, S \rangle = \text{argmax} \left(\sum_{\hat{s} \in \hat{S}, s \in S} \left(\sum_{m \in M} m(\hat{s}, s) \right) \right) \quad (1)$$

Here, S and $\hat{S}$ denote the reference and generated summaries, respectively, and M represents the set of evaluation metrics.

3.2 Human Evaluation Metrics

Human evaluation metrics offer a reference-free perspective, assessing the quality of the dataset and baselines using the source article. Evaluations focus on two aspects: overall summarization quality and aspect quality.

Following Fabbri et al. (2021), we assess summaries using four criteria: (1) *Coherence*, evaluating the overall coherence of the summary; (2) *Consistency*, evaluating factual alignment of the summary with the source; (3) *Relevance*, focusing on the inclusion of key content from the source; and (4) *Fluency*, examining the linguistic quality of the summary.

Furthermore, in the context of DABS, an additional metric, *Aspect-Quality*, is introduced. This metric evaluates the distinctiveness and focus of aspect-based summaries, ensuring that each aspect is clearly represented and remains the central focus within its respective summary (Angelidis et al., 2021; Amplayo et al., 2021).

4 Datasets

The Disordered-DABS dataset comprises two subsets: **D-CnnDM** and **D-WikiHow**, each adapted from existing summarization datasets. Instead of constructing these subsets from the ground up, we opted for an automatic conversion approach, resulting in large-scale samples that are well-suited for fine-tuning purposes.

4.1 Dataset Creation

Each sub-dataset is created using distinct methods to generate dynamic aspect-based article-summary pairs from original article-summary pairs.

D-CnnDM derivedfrom the CNN/DailyMail dataset (See et al., 2017), following the aggregation approach of Frermann and Klementiev (2019). Each sample in D-CnnDM consists of up to 11 randomly selected instances from CNN/DailyMail,

Dataset	Domain	#Train	#Valid	#Test	Avg.Arc. Len	Avg.Sum. Len	Avg.# Asp
D-CnnDM	News	47,920	2,185	1,917	4,143 (233)	311 (172)	6.00 (3.15)
D-WikiHow	Encyclopedia	142,284	20,327	40,653	452 (482)	41 (34)	6.23 (4.70)

Table 2: Statistics of the datasets. "Avg.Arc.Len", "Avg.Sum.Len", and "Avg.#Asp" denote the average length of articles, summaries, and the average number of aspects per sample. Standard deviations are provided in brackets.

treating each instance as a distinct aspect, thereby providing coarse-grained aspects.

D-WikiHow modifies the WikiHow dataset (Koupaee and Wang, 2018), where each article is composed of multiple paragraphs, each introduced by a bold summary sentence. In D-WikiHow, each summary sentence is treated as an individual aspect.

After aggregation, sentences within source articles are intentionally shuffled to produce disordered texts. This design aims to challenge models to cluster sentences by their aspects and generate precise aspect-based summaries. Such an approach not only tests the models' proficiency with ordered texts but also enhances their applicability to scenarios involving disordered texts.

An overview of the datasets is provided in Table 2, highlighting the significant variability in article lengths, summary lengths, and aspect numbers. More detailed distribution statistics are available in Appendix A. The samples of the dataset are shown in Appendix A.2

4.2 Dataset Quality

We conducted a human evaluation of the datasets, adhering to the methodology described in Section 3.2. For this assessment, we randomly selected thirty samples from each of D-CnnDM and D-WikiHow. Each sample was rated by three annotators on a scale from 1 to 5 across five distinct metrics. The specifics of the survey design are elaborated in Appendix C.

The results, presented in Table 3, confirm the high quality of both datasets. Notably, the scores for "Coherence" and "Fluency" are impressive, a reflection of the human-crafted nature of the source summaries. D-CnnDM outperforms D-WikiHow in terms of "Consistency" and "Relevance", which could be attributed to inherent limitations in the original WikiHow dataset as it is observed that D-WikiHow summaries sometimes either lack necessary details or include extraneous information.

Regarding "Aspect Quality", D-CnnDM showcases more clearly defined and distinct aspects compared to D-WikiHow. The aspects in D-CnnDM were noted by the annotators for their clarity, whereas D-WikiHow was marked by a certain vagueness in aspect demarcation, leading to overlaps in the content of summaries. This variance is likely due to the differing methodologies used in aspect generation for each dataset. Annotators also noted that their ratings for "Aspect Quality" were relatively more objective than other criteria. However, they expressed some reservations about the certainty of these scores, indicating the inherent subjectivity in such evaluations.

	Coherence	Consistency	Fluency	Relevance	Aspect
CnnDM	4.8 (0.4)	4.8 (0.4)	5.0 (0.2)	4.9 (0.3)	4.8 (0.8)
WikiHow	4.3 (0.9)	3.7 (1.2)	4.2 (1.0)	3.8 (1.3)	3.9 (1.3)

Table 3: Average (std) human evaluation ratings (1–5 scale) on the five quality criteria, determined by 30 instances from the test samples. "Aspect": Aspect-Quality, "CnnDM":D-CnnDM and "WikiHow":D-WikiHow

We also assessed the effect of sentence shuffling in the datasets. According to annotator feedback, the disordered texts in D-CnnDM did not substantially hinder the identification of topics or aspects. In contrast, D-WikiHow presented greater challenges in comprehending aspects when sentences were shuffled, exacerbated by the pre-existing ambiguity of the aspects. Thus, for human annotators, at least, disordered textual environments increase the complexity of the task, particularly when aspects or topics are not clearly defined.

5 Baseline Models

In this section, we outline the baseline approaches developed to address the unique challenges presented by our Disordered-DABS dataset, particularly focusing on aggregating aspect-based information from disordered texts. We explore two primary strategies: (1) Clustering-Based-Summarization (*Cluster*), which involves clustering sentences by aspects before summarization, and (2) Keyword-Guided (*Keyword*), utilizing aspect-related keywords to guide the summarization pro-

cess. Additionally, the Prompt-Guided Summarization (*Prompting*) is examined, leveraging large-language models for aggregating aspect-based information and generating summaries.

Both the *Cluster* and *Keyword* methods employ BERTopic (Egger and Yu, 2022) for sentence clustering and keyword generation, taking advantage of its capabilities in unsupervised clustering and topic modeling. The hyperparameters of BERTopic models are fine-tuned to align the number of references and generated summaries in validation datasets, with further details provided in Appendix B.4. Concurrently, the *Prompting* method utilizes prompts to direct large-language models, such as GPT-3.5, in automatically aggregating aspect-based information and crafting summaries. Based on a previous study (Guo and Vosoughi, 2023), we control the length of the generated summaries of all models to mitigate the influence of the generated summary length.

5.1 Clustering-Based-Summarization

The *Cluster* method, inspired by previous work (Hayashi et al., 2021; Amar et al., 2023), employs BERTopic to cluster sentences based on their respective aspects. Unlike their supervised approach, we utilize an unsupervised model within BERTopic to handle dynamic aspects with varying numbers and contents. Post-clustering, an abstractive summarization model based on Longformer-Encode-Decoder (Beltagy et al., 2020), is used for generating summaries. This summarization model is fine-tuned using original training sets formatted as one-document-to-one-summary, which may potentially inflate its performance.

5.2 Keyword-Guided Summarization

The *Keyword* approach, drawing on insights from previous studies in ABS and DABS (Ahuja et al., 2022; He et al., 2022; Yang et al., 2023b), directs a conventional summarization model to generate aspect-based summaries. This is achieved by incorporating keywords related to each aspect of the source article. For the extraction of unsupervised keywords, we employ C-TF-IDF from BERTopic (Grootendorst, 2022), identifying the ten most pertinent keywords for each aspect. These keywords, denoted as K, are then prefixed to the shuffled source document D, with a "[SEP]" token separating them (for example, K [SEP] D). It's important to note that the number of aspects must be specified in BERTopic during the training phase to facilitate keyword generation. For the *Keyword* method, we leverage the Longformer-Encode-Decoder model (Beltagy et al., 2020) as it demonstrates superior performance in previous DABS benchmarks and can accommodate inputs up to 16k characters.

5.3 Prompt-Guided Summarization

The *Prompting* method employs GPT-3.5[3], a leading large language model, for the Disordered-DABS task. Given its proven efficacy in ABS tasks with pre-defined aspects (Yang et al., 2023a), we propose that well-designed prompts can effectively direct GPT-3.5 in collating aspect-based information and producing summaries. The specifics of these prompts are outlined in Appendix B.3. To circumvent the context length limitations of GPT-3.5, we utilize different variants of the model suited to each dataset: GPT-3.5-turbo-16k for D-CnnDM and GPT-3.5-turbo for D-WikiHow, selected for their best performance within the respective context length boundaries.

6 Evaluation and Results

We assessed the performance of our baseline models, as detailed in Section 5, employing both automatic (Section 3.1) and human evaluation methods (Section 3.2). Due to the context length constraints of these baseline models, we truncated both the source articles and their corresponding aspect-specific summaries according to their average lengths and standard deviations. Summaries that did not incorporate any sentences from their source articles were omitted from the reference set.

While fine-tuning our model, we imposed a maximum limit on the number of aspects. However, this limit was not enforced during the evaluation phase. Consequently, baseline models could face penalties for inaccuracies in predicting the correct number of aspects. We have provided additional information about the data preprocessing specific to each dataset in Appendix B.2.

6.1 Automatic Evaluation Results

The automatic evaluation encompassed both summary quality and aspect identification accuracy. For *Keyword* and *Cluster*, multiple experiments were conducted using varied random seeds. For *Prompting*, a small-scale study identified the most effective prompt, with details in Appendix B.5.

[3]Our choice of GPT-3.5 over GPT-4 was influenced by budget constraints and the latter's context length limit of 4k tokens

Model	Dataset	#AbsAspDiff	BERTScore	Rouge-1	Rouge-2	Rouge-L
Keyword		1.3 (0.0)	14.2 (0.1)	24.8 (0.0)	8.9 (0.2)	17.2 (0.1)
Cluster	D-CnnDM	1.3 (0.0)	15.1 (0.2)	25.4 (0.1)	9.1 (0.1)	17.1 (0.1)
Prompting		6.2	9.1	12.4	4.1	9.1
Keyword		2.7 (0.1)	30.5 (0.1)	14.6 (0.2)	5.0 (0.1)	14.2 (0.2)
Cluster	D-WikiHow	2.7 (0.1)	31.3 (0.0)	19.1 (0.1)	7.8 (0.1)	18.5 (0.1)
Prompting		5.5	17.7	11.4	3.8	10.3

Table 4: The performance of baselines across D-CnnDM and D-WikiHow. Mean scores are reported, accompanied by standard deviations in brackets. Due to budgetary constraints, the results for GPT-3.5 are derived from a single experimental run.

The results, shown in Table 4, indicate that the two-step methods (*Keyword* and *Cluster*) generally outperform *Prompting* across all metrics, particularly in #AbsAspDiff. This trend suggests that *Prompting* might struggle to accurately discern the nuances of different aspects within samples. Notably, *Cluster* exhibits superior performance over *Keyword*, suggesting that simply providing automatic generation of aspect-related keywords is insufficient for effectively extracting aspect-based information from disordered texts.

A pilot check was conducted to understand the disparity in aspect identification between *Prompting* and the other baselines. This check reveals that *Prompting* frequently splits a single topic or aspect into multiple segments, especially in D-CnnDM samples with one or two topics. This indicates that *Prompting*'s zero-shot in-context learning approach may have difficulties in recognizing the cohesion of a single topic or aspect. Conversely, *Cluster* and *Keyword*, leveraging BERTopic for aspect identification, appear to have a more accurate grasp of topic or aspect content.

6.2 Human Evaluations

In our study, human evaluations complemented automatic metrics by assessing the performance of models on the D-CnnDM and D-WikiHow subsets. We selected thirty instances from each dataset for evaluation by three judges, based on the criteria outlined in Section 3.2 and using the interfaces described in Appendix C, with results detailed in Table 5.

The evaluations revealed diverse performances across datasets and evaluation criteria. In the D-CnnDM dataset, *Prompting* excelled in two specific criteria, whereas *Cluster* was superior in three. For the D-WikiHow dataset, *Prompting* outperformed all other baselines across all criteria, indicating its greater effectiveness in less aspect-distinct scenarios. Consistent with automatic evaluations, *Cluster* generally surpassed *Keyword* in performance.

Notably, *Prompting* was consistently rated highest for "Fluency" and Aspect Quality." This is in contrast to automatic evaluations, where *Prompting* was penalized for generating fewer aspects. Further inquiry with judges revealed that *Prompting* tends to fragment one aspect into several, accounting for this discrepancy. Additionally, annotators' preference for a higher number of aspects in D-WikiHow, as opposed to reference standards, highlights the subjective nature of determining the optimal number of aspects, especially in nuanced aspect differentiation.

6.3 Expmale with Different Baselines

Figure 2 presents an illustrative example featuring the source article, reference summaries, and summaries generated by all baselines. The source article addresses the topic of sending a friend invitation on Facebook. We have pre-aligned the reference and generated summaries for ease of comparison. The presence of empty quotes ("") signifies the absence of a corresponding generated summary for a given reference summary. Notably, *Prompting* generates six aspect-based summaries while the other two only generate one aspect-based summary with information missing. The difference in aspect discovery preference is in alignment with what we observe in other experiments.

7 Ablation Analysis

Our ablation study delves into three key areas: 1) Exploring the few-shot learning capabilities of GPT-3.5; 2) Assessing the impact of disordered text settings on model performance; and 3) Exam-

Dataset	Model	Coherence	Consistency	Fluency	Relevance	Aspect-Quality	Rank
D-CnnDM	*Keyword*	3.17 (0.76)	2.79 (0.93)	3.38 (0.97)	3.25 (0.68)	3.00 (1.14)	2.21 (0.83)
	Cluster	3.50 (0.72)	3.21 (0.78)	3.50 (0.72)	3.46 (0.88)	4.04 (1.00)	1.71 (0.69)
	Prompting	3.21 (1.28)	2.71 (0.75)	4.29 (1.12)	3.12 (0.80)	4.25 (1.07)	2.08 (0.88)
D-WikiHow	*Keyword*	2.16 (1.29)	2.58 (0.72)	2.48 (1.03)	1.90 (0.94)	1.97 (1.35)	2.68 (0.48)
	Cluster	2.74 (1.18)	3.23 (0.99)	3.06 (1.03)	2.84 (1.00)	2.55 (1.29)	2.32 (0.73)
	Prompting	4.55 (0.57)	3.58 (1.09)	4.39 (0.72)	4.10 (1.14)	4.58 (0.81)	1.00 (0.00)

Table 5: Average (std) human evaluation ratings (1–5 scales) on the five quality criteria and the ranking (1–4 scales), determined by 30 instances from the test samples.

Source Article	"Once the reset process is complete, your phone will be formatted as it was when it left the factory. It's often a gear-shaped icon (⚙) or an icon containing a series of slider bars.. It's in the "Personal" section of the menu.. On Samsung devices, first tap General management and then tap Reset.. On Samsung devices the option will say Reset Device.. You may be asked to enter your device's passcode depending on your security settings.."
Reference	"Open your device's Settings. " "Scroll down and tap Backup & reset. " "Tap Reset Phone. "
Keywords	"Open the Settings app." "" ""
Cluster	"" "" "Tap Reset Device."
GPT-3.5	"You might need to enter your device's passcode depending on your security settings." "For Samsung devices, go to General management and then tap Reset." "On Samsung devices, the option will be labeled Reset Device." "The reset icon is usually a gear-shaped or slider bars icon." "Find the reset option in the "Personal" section of the menu." "Your phone will be reset to its factory settings after the reset process"

Figure 2: An example of the source article, reference summaries, and generated summaries by baselines. Empty quotes (“ ”) indicate missing generated summaries.

ining the effects of variations and discrepancies in aspect number.

7.1 Few-shot Settings for GPT-3.5

The results for *Prompting*, as detailed in Table 4, were obtained using zero-shot learning. To further assess the capabilities of few-shot in-context learning and fine-tuning, we conducted additional experiments with the D-WikiHow dataset. These experiments were not extended to the D-CnnDM dataset due to the constraints of GPT-3.5-turbo-16k, specifically its limitations on fine-tuning and the context length restrictions applicable to few-shot in-context learning.

Table 6 shows that few-shot learning and fine-tuning can enhance the performance of GPT-3.5 by minimizing the absolute differences in aspect counts between the predicted and ground truth summaries [4]. This implies that GPT-3.5 still requires domain-specific information to accurately differentiate between various aspects or topics. Interestingly, we observe that few-shot learning outper-

[4]Samples are shown in Appendix D

Model	#Sample	#AbsAspDiff	BERTScore	Rouge-1	Rouge-2	Rouge-L
zero-shot	0	5.6	18.3	11.7	4.0	10.6
In-context	1	4.1	26.6	16.6	5.9	15.5
	3	3.7	29.0	18.1	6.7	17.0
	6	3.5	29.6	18.6	7.0	17.5
Fine-tuning	50	4.2	17.1	11.0	4.3	10.2
	100	3.4	25.4	16.1	6.7	15.3
	200	4.6	15.1	10.1	4.1	9.4

Table 6: Comparison of zero-shot learning, in-context learning, and fine-tuned *Prompting* based on sampled data. "#Samples": the number of samples used.

Model	Dataset	#AbsAspDiff	BERTScore	Rouge-1	Rouge-2	Rouge-L
Keyword	D-CnnDM	1.2 (0.0)	2.0 (-12.2)	18.7 (-6.1)	5.3 (-3.6)	12.7 (-4.5)
Cluster		1.2 (0.0)	19.0 (2.8)	29.0 (2.6)	12.4 (2.3)	20.5 (2.4)
Prompting		5.6 (-0.6)	12.3 (3.2)	16.9 (4.6)	6.0 (1.9)	11.9 (2.8)
Keyword	D-WikiHow	2.7 (0.0)	27.5 (-3.0)	15.9 (1.3)	6.0 (1.0)	15.3 (1.1)
Cluster		2.7 (0.0)	29.9 (-1.4)	18.2 (-0.9)	7.6 (-0.2)	17.6 (-0.9)
Prompting		4.8 (-0.7)	20.5 (2.8)	13.3 (1.9)	4.7 (0.9)	12.1 (1.8)

Table 7: The performance of baseline models on the D-CnnDM and D-WikiHow sub-datasets, noting the effects of organized settings in parentheses. For the "#AbsAspDiff" metric, a negative value signifies an improvement, whereas the opposite is true for the remaining four metrics.

forms fine-tuning when the sample size is limited, suggesting that it may be a more cost-effective approach under similar conditions.

7.2 The Impact of Disorder Texts

Here, we investigate how disordered text environments, a central aspect of the Disordered-DABS dataset, affect summarization model performance. By conducting control experiments with organized texts under identical experimental setups as previously outlined in Section 6.1, the study reveals varied impacts on model efficacy across different methods and datasets (see Table 7).

Specifically, the *Prompting* approach showed uniform improvement in performance metrics for both the D-CnnDM and D-WikiHow datasets when applied to organized texts. In contrast, the *Keyword* and *Cluster* approaches demonstrated mixed results, with the former seeing a decline in performance on the D-CnnDM dataset but an improvement on D-WikiHow (with an exception for the #AbsAspDiff metric), while the latter improved on D-CnnDM but deteriorated on D-WikiHow. These outcomes underscore the complex nature of disordered texts' impact on summarization tasks, suggesting the need for deeper exploration into how text structure affects summarization model performance.

7.3 The Influence of Aspect Number Variations and Discrepancies

Given the significance of aspect count in Disordered-DABS, we investigate how varying aspect numbers impact the performance of the models.

Figure 3 demonstrates how the performance of various baseline models is affected by changes in the number of aspects across two datasets. In the D-CnnDM dataset, *Prompting* improves as the number of aspects increases, whereas *Keyword* and *Cluster* show a decline in performance. Conversely, in the D-WikiHow dataset, the number of aspects has a negligible effect on *Prompting* but adversely affects *Keyword*. These observations highlight that the impact of aspect numbers varies significantly across different models and datasets.

To further assess if our model's improved performance is due to its ability to accurately predict the number of aspects or the quality of the generated summaries, we experimented by matching the ref-

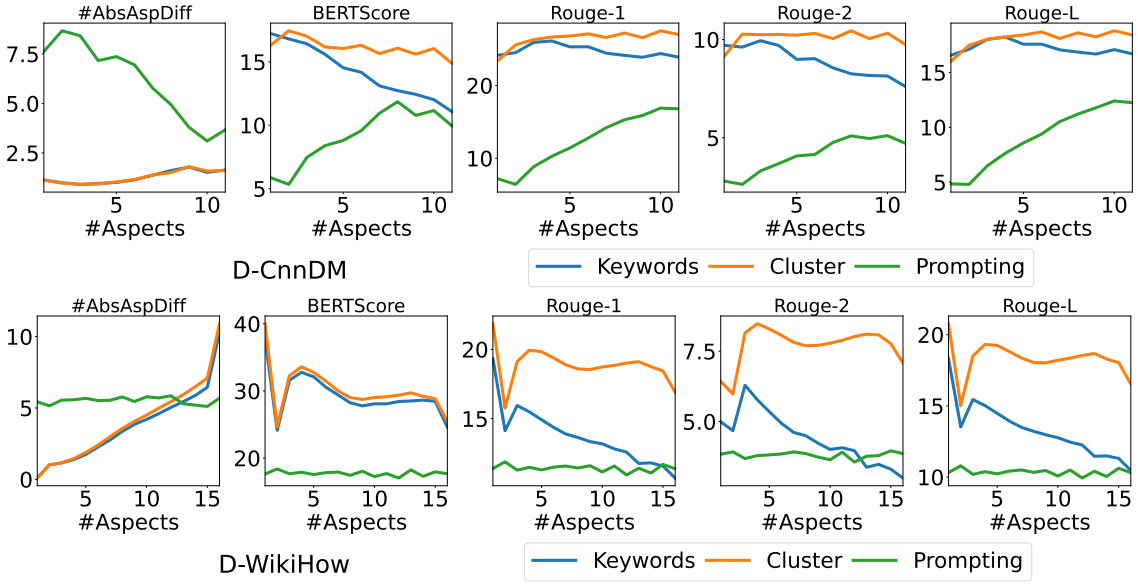

Figure 3: Performance variation of baselines with changes in aspect number in the reference. The final data point represents cases where the number of aspects in the reference summary equals or exceeds 12 for D-CnnDM and 16 for D-WikiHow.

Model	Dataset	BS	R1	R2	RL
Keyword	D-CnnDM	19.4(5)	32.1(7)	11.9(3)	22.4(5)
Cluster		20.5(5)	33.0(8)	12.4(3)	22.4(5)
Prompting		21.1(12)	26.4(14)	9.7(6)	19.5(10)
Keyword	D-WikiHow	51.5(21)	29.8(15)	11.9(7)	29.0(15)
Cluster		55.0(24)	38.1(19)	18.2(10)	37.0(19)
Prompting		35.5(18)	26.4(5)	12.5(9)	24.3(14)

Table 8: The performance of all baselines across datasets with "perfect-match". Mean scores are reported, accompanied by the improvement caused by "perfect-match" settings in brackets. "BS":"BERTScore","R1":"Rouge-1","R2":"Rouge-2","RL":"Rouge-L".

erence and generated summary numbers ("perfect-match"). For each sample, we consider only the top-k aspects where each model achieves optimal performance, and k is the smallest aspect number among all models and the reference. The results, displayed in Table 8, show significant improvement for all baselines, particularly for D-WikiHow. This suggests that identifying and combining aspects in D-WikiHow presents more challenges than in D-CnnDM, likely due to the methodologies employed for aspect generation and the inherent ambiguity in defining aspects.

8 Conclusion

Our Disordered-DABS benchmark introduces the first dataset dedicated to dynamic aspect-based summarization in disordered texts, innovatively crafted from high-quality summaries of existing datasets. This approach sidesteps the conventional challenge of manual summary generation by transforming available datasets for summarization.

The evaluation of leading-edge baseline models has exposed a notable deficiency in the current capacity to effectively tackle this complex task. Our analysis shows that contemporary models struggle significantly with the nuances of dynamic aspect-based summarization in disorganized texts, underscoring the task's complexity and highlighting avenues for future research and enhancement.

Limitations

While not raising significant ethical concerns, our study acknowledges limitations associated with the use of publicly available datasets, which may contain sensitive or potentially offensive content. We recommend caution in handling these datasets. The transformation of existing summaries to create dynamic aspect-based summaries in disordered texts, despite manual checks, may not ensure uniform quality across all samples, prompting users to verify anomalies in further experiments.

The Disordered-DABS benchmark, derived from two distinct datasets, offers variety in domains and granularity but does not cover the entire spectrum of dynamic aspect-based summarization challenges in disordered texts. It serves as a baseline for general domains, with a recommendation to adapt our methodology for domain-specific benchmarks in dynamic aspect-based summarization.

Furthermore, our reliance on GPT-3.5-turbo introduces potential biases due to undisclosed training data, possibly including our test data, which could influence our findings. The prompt-dependent nature of our experiments with GPT-3.5-turbo (the specific prompts utilized are detailed in the Appendix) underscores that our conclusions about its effectiveness are contingent on the specific prompts used, indicating that different prompts could yield varied results. This highlights the need for transparent model documentation and the exploration of prompt-independent evaluation methods to enhance the reliability of findings in dynamic aspect-based summarization research.

Acknowledgements

This work is supported in part by NSF Award 2242072 and a generous grant by the Templeton Foundation.

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. Aspectnews: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. Openasp: A benchmark for multi-document open aspect-based summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1991.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7.

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.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Xiaobo Guo and Soroush Vosoughi. 2023. Length does matter: Summary length can bias summarization metrics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15869–15879.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

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

Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *Proceedings of the 18th international conference on World wide web*, pages 131–140.

Mounica Maddela, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2022. Entsum: A data set for entity-centric extractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3355–3366.

Dhruv Mehra, Lingjue Xie, Ella Hofmann-Coyle, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2023. Entsumv2: Dataset, models and evaluation for more abstractive entity-centric summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5538–5547.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1080–1089.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel Bowman. 2022. Squality: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156.

Yumo Xu and Mirella Lapata. 2021. Generating query focused summaries from query-free resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6096–6109.

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

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023b. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022. Unsupervised multi-granularity summarization. *arXiv preprint arXiv:2201.12502*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, 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.

Dataset	Domain	#Train	#Valid	#Test	Avg.Arc. Len	Avg.Sum. Len	Avg.# Asp
AnyAspect	News	287,227	13,368	11,490	681 (332)	20 (18)	7.06 (4.20)
OASUM	Wikipedia	1,937,776	60,526	60,481	792 (1,181)	35(37)	1.8 (1.48)
OpenASP	News	476	238	596	7,930	96	3.1
ENTSUMV2	News	N/A	N/A	N/A	1,002	46	N/A

Table A1: Statistics of the datasets. "Avg.Arc.Len", "Avg.Sum.Len", and "Avg.#Asp" denote the average number of words of articles, summaries, and the average number of aspects per sample. Standard deviations are provided in brackets. Since the dataset ENTSUMV2 is not publicly available, we only report the statistics in their paper.

A Dataset

A.1 Disordered-DABS Distribution Statistics

In this appendix, we analyze the distribution of various statistics in our datasets, as shown in Figure A1. For D-CnnDM, the article lengths approximate a normal distribution, while the summary lengths follow a long-tail distribution. In contrast, for D-WikiHow, both article and summary lengths exhibit long-tail distributions.

The aspect number distribution in D-WikiHow shows a right-skewed pattern, whereas, in D-CnnDM, it is almost uniform. This difference allows us to investigate how aspect number distribution influences dynamic aspect-based summarization. Given these variations, D-CnnDM may offer a wider variety of aspects due to its balanced distribution.

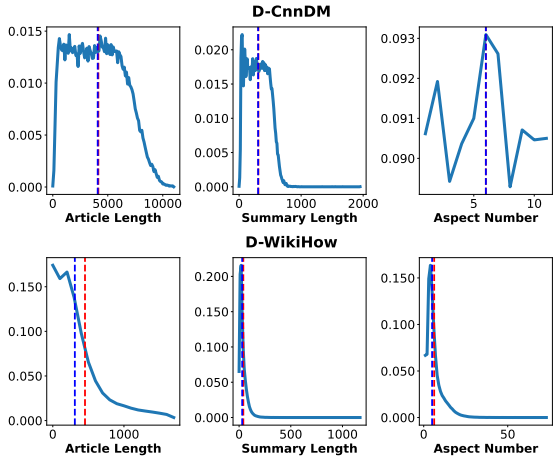

Figure A1: Distributions of article length, summary length, and aspect number across datasets. Mean and median values are marked with red and blue lines, respectively. For clarity, distributions are shown within three standard deviations from the mean for source articles and references.

A.2 Disordered-DABS Samples

Figure A2 and Figure A3 show the source articles and aspect-based summaries, which belong to 4 randomly selected different instances from Disordered-DABS (2 from D-CnnDM and two from D-WikiHow).

A.3 Details on DABS Datasets

Table A1 provides detailed information about the datasets previously introduced in Table 1.

B Experimental Details

B.1 Computing Infrastructure

For our experiments, we used a Lambda machine equipped with 250 GB of memory, 4 RTX 8000 GPUs, and 32 CPU cores. The machine runs on Ubuntu 20.04, and our experiments are conducted using Python 3.8.10. The CUDA version is 11.9, and the GPU Driver Version is 520.61. The main packages we utilize include bertopic (0.14.1), cuml-cu11 (23.4.1), deepspeed (0.8.0), torch (1.13.1), scikit-learn (1.1.2), sentence-transformers (2.2.2), scipy (1.9.1), transformers (4.22.1), and numpy (1.23.3). The complete list of packages will be provided in the code release.

B.2 Data Processing

Table B2 displays the thresholds for source article length, single-aspect summary length, and the maximum aspect number used for truncation during the experiments. The threshold for the summary length corresponds to the single-aspect length, so the total length of the summaries is multiplied by the aspect number.

For D-CnnDM and D-WikiHow, we truncate the source article and single-aspect summaries based on the mean and standard deviation of the length (approximately equal to mean + 2 $\times$ std).

Regarding the maximum aspect number, we set the thresholds for D-CnnDM based on the largest

Source Article	E-cigs contain far fewer cancer-causing and other toxic substances than cigarettes, however their long-term effects on health and nicotine dependence are unknown. The airstrikes prevented the bombing, the military said. Cigarette-smoking teenagers hit a new low last year as youngsters turn to e-cigarettes instead, a study has revealed. There are currently more than 400 brands of 'e-cigs' available. The Israel Defense Forces said the airstrikes were carried out in the occupied Golan Heights against four militants who crossed into the area from Syria. 'What's most surprising is how incredibly rapid the use of products other than cigarettes has increased,' Frieden said in an interview quoted by The Washington Post. It is home to 41,000 residents, including Jews, Druze and Alawites. They have been on sale in the United States since 2006, but this is the first year that the Centers for Disease Control has measured their use. 'The concern is that e-cigarette advertising is recruiting intermediate risk adolescents to nicotine use- kids who would not otherwise have started smoking,' said James Sargent, of Dartmouth Hitchcock's Norris Cotton Cancer Center in a paper published in the journal Pediatrics. The CDC study also revealed that about nine per cent of teenagers surveyed said they had tried hookahs. Tom Frieden, CDC director, remarked that the results are 'alarming' and 'shocking'. Around two million high school students admitted to buying vaporizers in 2014 - more than triple the 660,000 recorded the year before. It follows a government-funded study released in December which found similar trends, leading experts to warn of a potential 'epidemic of teen tobacco use'. Jerusalem (CNN)The Israeli military conducted airstrikes Sunday night in the area between Israel and Syria, targeting a group of militants allegedly trying to plant a bomb on the Israeli border. There are currently more than 400 brands of 'e-cigs' available in the US. Researchers have been critical of the fact that packaging and marketing of e-cigarettes is not regulated, and that manufacturers' opportunities to target adolescents are wide open, absent FDA regulation of these products. It was not immediately known to what militant group the men belonged. On the rise: Around two million high school students now admit to buying e-cigarettes . 'It is subjecting another generation of our children to an addictive substance.' The Golan is regarded internationally as occupied territory despite Israeli governmental control. E-cigs contain far fewer cancer-causing and other toxic substances than cigarettes, however their long-term effects on health and nicotine dependence are unknown. The popularity of e-cigarettes, which typically deliver nicotine, propylene glycol, glycerin and flavorings through inhaled vapor, has increased in the past five years. "A group of armed terrorists approached the border with an explosive device, which was intended to be detonated against IDF forces," the Israeli military said. But smoking of traditional cigarettes plummeted to about nine per cent. Israel seized the territory from Syria during the 1967 Israel-Arab war, and it was eventually annexed. Three of the alleged attackers were killed, Israeli media reported, citing IDF sources. The CDC report, released on Thursday, is based on a national survey of about 22,000 students at middle schools and high schools, both public and private. 'If this pattern of use is adopted by adolescents in the continental U.S., we could be in for an epidemic of teen tobacco use in this country that could greatly reduce the overall benefits to public health of e-cigarettes.' E-cigarettes typically deliver nicotine, propylene glycol, glycerin and flavorings through inhaled vapor. 'These are kids who might go on to smoke cigarettes, which are much better at delivering nicotine than e-cigarettes.
Aspect-based Summary	The Israeli military says the militants were trying to plant a bomb . The men crossed from Syria, Israel says . Two million high school students admitted to using e-cigarettes in 2014 . That is an increase of 13 per cent from the 660,000 recorded in 2013 . Traditional cigarette use has plummeted by 9 per cent, CDC study shows . Experts warn e-cigarette marketing is not regulated .

Source Article	Anyone who recognises the collection, or who has any information about where it may have come from, is asked to contact DS Lewis on 101. Kadyrov appeared to defend the suspect Zaur Dadayev by calling him a 'true patriot' and a 'deeply religious man'. Andrey Lugovoy (pictured), 48, was given a medal for 'services to the motherland' as the UK holds a public inquiry into the 2006 killing of Mr Litvinenko . 'We finished about seven in the morning. The Russians allege the West was complicit in the toppling of pro-Moscow Yanukovych so threatening Russia's national security. Alternatively, call Crimestoppers anonymously on 0800 555111. The £50,000 find was made in London Colney last month and detectives believe the bullion may have been stolen in a burglary. The items were mostly found individually wrapped in cellophane or were held in plastic containers. Andrey Lugovoy, 48, was given a medal for 'services to the motherland' as the UK holds a public inquiry into the 2006 killing of Mr Litvinenko. The box, containing the coins, doubloons and bullion, was discovered under a hedge n Hertfordshire last month . Vladimir Putin yesterday honoured the man Scotland Yard believes poisoned dissident Alexander Litvinenko with polonium in London. 'I invited the leaders of our special services and the defence ministry to the Kremlin and set them the task of saving the life of the president of Ukraine, who would simply have been liquidated,' said Putin. Kadyrov received his state award for 'work achievements, strenuous social activities and long conscientious service', but it comes amid claims that even some senior Russian officials are worried about the sway held by the Chechen. In a trailer for a new TV documentary, he admitted calling his security chiefs to the Kremlin for a secret meeting to order them to save the life of deposed Ukrainian president Viktor Yanukovych. Putin had blocked the extradition of Lugovoy, a former secret-service operative who is now an MP with the ultra-nationalist Liberal Democratic Party. Hidden treasure: This silver bullion was among a stash of valuable items handed into Hertfordshire Police . 'The items are very distinctive so we hope that someone might recognise them and that they can be returned.' He is chief suspect in the murder of former secret agent Mr Litvinenko, who from his death bed accused Putin of ordering his killing. Lugovoy holds the post of deputy chairman of the Russian parliament's security and anti-corruption committee. Also found in the stash, valued at £50,000, were a collection of gold and silver coins and doubloons . Opposition politician Ilya Yashin said Putin was sending a signal: 'These are my people, do not touch them.' Putin revealed yesterday that he held an all-night conclave of his defence and secret services cardinals ahead of his order last year to restore Crimea to Russia. Vladimir Putin honoured the man Scotland Yard believes poisoned dissident Alexander Litvinenko (pictured) with polonium in London . Police are searching for the owners of a collection of gold and silver bullion, coins and doubloons which were discovered under a hedge in Hertfordshire. The Russian president also gave the Order of Honour to Ramzan Kadyrov, the leader of Chechnya. Detective Sergeant Karen Lewis, from the St Albans Local Unit, said: 'Despite several enquiries, we have so far been unable to trace the owners of these items. The award came as Kadyrov admitted knowing the man who has confessed to killing Putin critic Boris Nemtsov in Moscow. When we were parting, I told all my colleagues, "We are forced to begin the work to bring Crimea back into Russia".'
Aspect-based Summary	Russia honours the man believed to have poisoned Alexander Litvinenko . Putin gave Andrey Lugovoy a medal for his 'services to the motherland' Lugovoy is believed to have poisoned dissident with polonium in London . Stash of treasure was found underneath hedge in Hertfordshire last month . Police believe the items were stolen in a burglary before being abandoned . So far, officers have been unable to locate the owner of £50,000 hoard .

Figure A2: Examples from the D-CnnDM, illustrating various aspects represented by different colors.

	Arc. Leng	Sum. Length	#Asp
D-CnnDM	11,264	76	12
D-WikiHow	2,040	20	16

Table B2: The thresholds for the source article length, single-aspect summary length, and the maximum aspect number for each sample.

$\#Asp$ in the datasets (12). For D-WikiHow, considering it follows a long-tail distribution, we set the threshold to 16 (approximately equal to the mean + 2 $\times$ std).

B.3 Prompts for GPT-3.5

The prompts employed for GPT experiments adhere to the following template: "You are a summarizer. Your task is to summarize sentences across multiple aspects or topics. The sentences are shuffled. Please generate multiple summaries, ensuring each summary pertains to a single aspect or topic. Limit each summary to one sentence, exclude the topic name, and restrict its length to fewer than N tokens. Separate each summary with [SEP]". Here, N represents the output length constraint. For zero-shot inference, the prompt is augmented with the phrase "There should be no more than M topics

Source Article	(It is not recommended for you to purposely arrive late at places.) If you already have dark hair, that's great! Be adventurous and daring! They're cheap and look kind of real.. If you can, try to craft a blue garter out of plastic to wear over the skirt. Due a certain malfunction of hers, she tends to run late a lot. A black pencil skirt is exactly what she wears, so go with that. (Nothing too crazy, unless you want to kill yourself.) For her rocket boots, try a pair of high heeled ones.. For the highlights though, if you don't want to dye your hair, try to buy some clip-ons. Robecca is into the steampunk look, more or less. Her favorite color is copper, but she also tends to wear a lot of blue. In order to gain her look, wear a blue tank top, with maybe a black cardigan jacket over it. Don't be afraid to try out a new stunt! Robecca is a daredevil, but very kind and caring, always trying to please others. Robecca has black curly hair, with blue highlights. Robecca is very limited on her makeup, but try for lavender eyeshadow and a light copper, almost like a caramel skin tone color, for the lipstick..
Aspect-based Summary	Get her attitude Dress like her. Makeup. Hair.

Source Article	The menu's icon looks like three horizontal lines. To ensure complete erasure of your online activity, it is recommended that you select these boxes as well.. These options will delete things other than your browsing history, such as your saved passwords, recent searches, and personal information. This will prompt a window to appear from which you will select the online activity that you wish to clear. Once you have specified what you want deleted, click on the button at the bottom of the window titled "Clear Browsing Data." You can do this by going to "History" and then "History" again in Chrome and simply "History" in Firefox.. If you wish to delete all of your internet history, select "The Beginning of Time" in Chrome or "Everything" in Firefox. Other options you may check include "Form and Search History," "Cookies," "Active Logins," and "Cache." In Firefox, this will be titled "Clear Browsing Data." After you open your browser, click on the menu in the upper right-hand corner. Your online activity should now be cleared.. From the dropdown menu, select how far back in time you would like to clear.
Aspect-based Summary	Open the menu. Go to your history. Select "Clear Browsing Data." Select what you wish to delete. Check the box entitled "Browsing History." Select "Clear Browsing Data."

Figure A3: Examples from the D-WikiHow, illustrating various aspects represented by different colors.

or aspects," serving to restrict the total number of aspects or topics considered. In this context, M denotes the maximum allowable count of topics or aspects.

B.4 BERTopic Hyperparameter Tuning

For the baselines, we performed a grid search to tune the BERTopic model hyperparameters. The hyperparameters tuned for the BERTopic include "n_neighbours", "n_component", and "min_dist", which control the cluster size and the samples within a cluster. We performed BERTopic clustering on each sample of the validation data and selected the combination of hyperparameters that minimized the absolute difference between the generated and reference aspect numbers.

B.5 Hyperparameters and Random Seeds

For our experiments, we used three random seeds (0, 10, and 42) for the complete dataset experiments. We used the Hugging Face implementation to fine-tune the language model with a batch size of 4 due to GPU memory limitations. The training epoch was set to 10 with an early stop, and all other training process hyperparameters were set to the default values provided by the package. We also used DeepSpeed to reduce memory requirements during the fine-tuning process. The specific DeepSpeed configuration will be provided along with our code.

C Human Annotation

The human annotation process is detailed in Figure C4, outlining a structured approach for annotators. They are instructed to thoroughly review the source article and its ground truth summary to: (1) Assess the ground truth based on five criteria: *Coherence*, *Consistency*, *Fluency*, *Relevance*, and *Aspect Quality*; (2) Evaluate generated summaries against these criteria, using the ground truth for reference; (3) Rank the summaries according to their overall effectiveness. All annotators possess at least a bachelor's degree and are proficient in English. They are compensated between $10 to $15 per hour for their contributions.

D Examples of Fine-tuned GPT-3.5

Figure D5 presents an illustrative example featuring the source article, reference summaries, and summaries generated by *Prompting* with the best performance of few-shot in-context learning and fine-tuning. We can observe that the generated summaries of *Prompting* are much better with few-shot inference/fine-tuning. However, we can also observe that both methods introduce extraneous information, potentially leading to hallucinations.

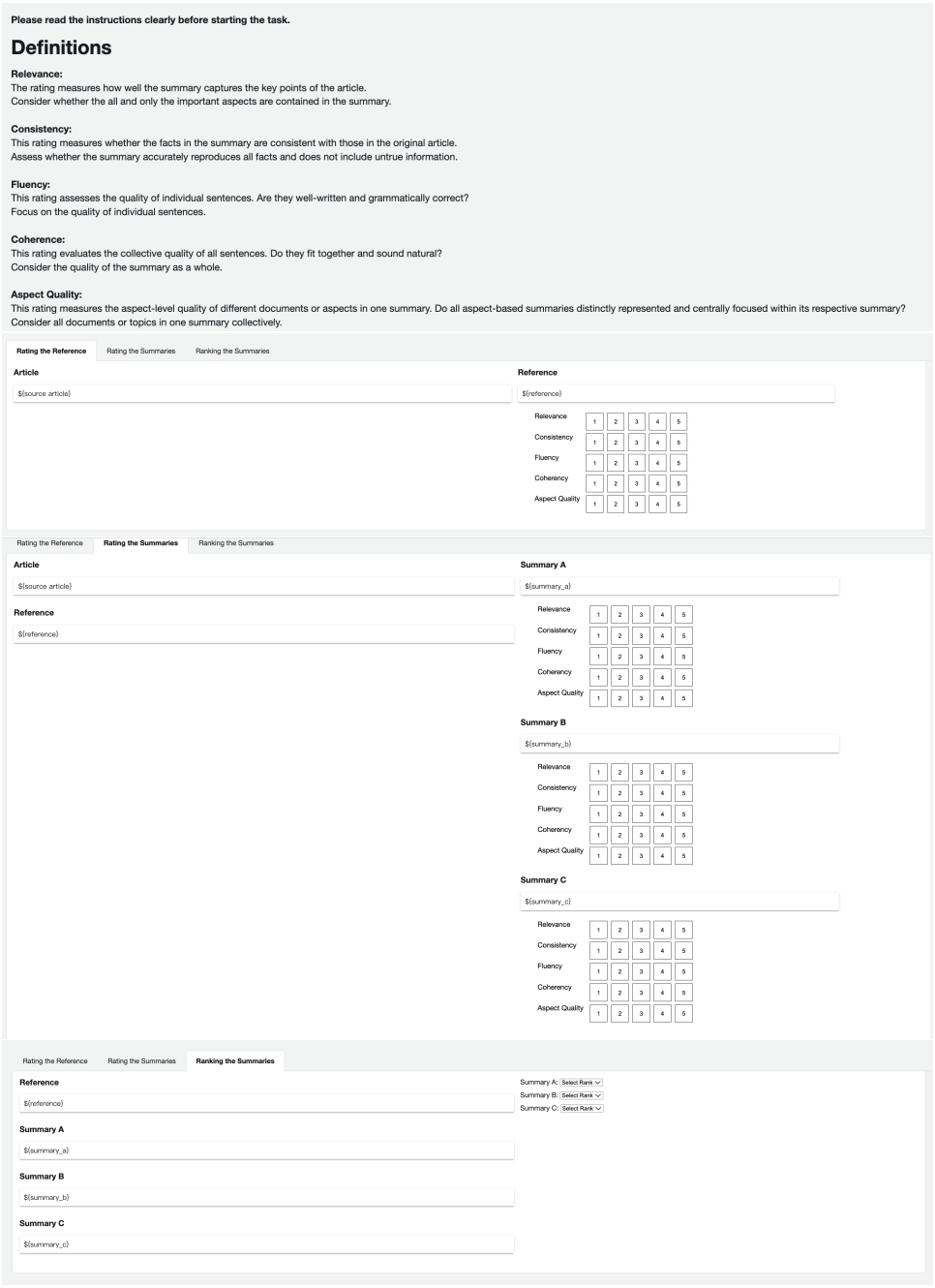

Figure C4: Example of the human annotation interface

Source Article	"Once the reset process is complete, your phone will be formatted as it was when it left the factory. It's often a gear-shaped icon (⚙) or an icon containing a series of slider bars.. It's in the "Personal" section of the menu.. On Samsung devices, first tap General management and then tap Reset.. On Samsung devices the option will say Reset Device.. You may be asked to enter your device's passcode depending on your security settings.."
Reference	"Open your device's Settings. " "Scroll down and tap Backup & reset. " "Tap Reset Phone. "
Zero-shot	"You might need to enter your device's passcode depending on your security settings." "For Samsung devices, go to General management and then tap Reset." "On Samsung devices, the option will be labeled Reset Device." "The reset icon is usually a gear-shaped or slider bars icon." "Find the reset option in the "Personal" section of the menu." "Your phone will be reset to its factory settings after the reset process"
6-shot-inference	"Access the settings menu on your device. " "Look for the "Reset" option. " "Enter your device's passcode if prompted. " "Select the option to reset your device. " "Follow the on-screen instructions to complete the reset process. "
100-sample Fine-tuning	"Open your phone's Settings app. " "Scroll down and tap on Backup and reset. " "Tap on Factory data reset. " "Tap on Reset phone. "

Figure D5: Example of the source article, reference summaries, and *Prompting* generated summaries with 6-shot inference and 100 samples fine-tuning.