Improving Aspect-Based Summarization via Contrastive Learning with Anchored Negative Examples

Elizabeth Palmieri **Yangfeng Ji**
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
{cxm7ja, yangfeng}@virginia.edu

Abstract

Text summarization helps users manage information overload, but traditional methods can be cumbersome when seeking specific details within a document. Aspect-based text summarization addresses this by using a query to guide which information should be summarized. However, distinguishing relevant from irrelevant information for a given aspect remains challenging in LLM-based summarization models. In this work, we propose utilizing contrastive learning to encourage LLMs to focus on aspect-related signals during training. We further design two variants of the learning algorithm, aspect-anchored and summary-anchored, corresponding to the strategies used in constructing negative examples. Evaluation with two representative LLM families (Llama 2 and Pythia) and two benchmark datasets (AnyAspect and CovidET) demonstrates the proposed methods' strong performance compared to their supervised fine-tuning and zero-shot counterparts, highlighting contrastive learning as a promising direction for aspect-based text summarization.[1]

1 Introduction

Aspect-based text summarization is a crucial task within Natural Language Processing that addresses the limitations of conventional summarization methods. Traditional models generate concise summaries of documents, aiming to save users time and effort. However, the varying lengths and diverse content of documents often render general summaries inadequate. Consider a news article covering an event from multiple perspectives – location, key figures, and the event itself. A user might be interested in a summary focused on the event's location rather than the individuals involved, as shown in Figure 1. Aspect-based summarization tackles this issue by providing both a document and a specific aspect, guiding the model to generate a summary tailored to that particular focus. The objective is to parse the document and selectively use the information only relevant to the given aspect.

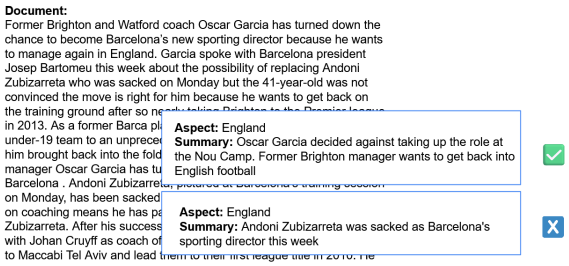


Figure 1: An illustration of aspect-based text summarization created based on an example from the AnyAspect dataset (Tan et al., 2020a). The first summary is related to the given aspect ENGLAND, while the second one is not.

A significant challenge in aspect-based text summarization lies in the model's ability to isolate and highlight aspect-specific information while effectively distinguishing it from the rest of the text. Without this crucial capability, the model risks generating a generic summary that fails to address the intended aspect. For example, a model presented with a news article about an earthquake (Ahuja et al., 2021) and tasked with generating summaries for two aspects (GEOGRAPHY and RECOVERY) might produce identical summaries detailing the earthquake's magnitude and recovery effort, neglecting the distinct nuances of each aspect. This underscores the need for a mechanism that enables the model to cluster similar information and separate dissimilar information within the latent space.

While prior work has established a foundation for understanding the capabilities of large language models (LLMs) in aspect-based text summarization (Mullick et al., 2024) and explored the impact

[1]The code and data will be available at https://github.com/elizabethpalmieri/contrastive-asp-summ.git

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 59–73

of contrastive elements (Gunel et al., 2023a), there remains a gap in informing aspect-based signals in LLMs. Furthermore, it lacks dedicated learning algorithms designed to teach LLMs to distinguish aspect-related signals during training (e.g., via comparing the aspects and their associated summaries). While adding instructions in prompts (e.g., "*Summarize this article with respect to Aspect within one short sentence.*") has shown promise (Yang et al., 2023a), a training-level contrastive loss function is necessary to effectively incorporate aspect information and emphasize crucial signals within the data.

To address this gap, we propose a novel approach that integrates contrastive learning into the fine-tuning of open-source LLMs to enhance aspect-based text summarization. We augment existing aspect-based summarization datasets to create two types of anchored negative examples: *aspect* anchored negative examples (same aspect with different summaries) and *summary* anchored negative examples (same summary with different aspects). During fine-tuning, the model is trained to differentiate either summaries or aspects with an anchor presented. By learning to discriminate between these pairs, the model gains additional context and generates summaries that are more closely aligned with the given aspect.

This work, to the best of our knowledge, is the first to integrate contrastive learning into LLM fine-tuning specifically for aspect-based text summarization. Our contributions are threefold: (1) We design a novel contrastive learning algorithm with anchored negative examples for LLMs fine-tuning on aspect-based text summarization; (2) We perform an empirical study evaluating the performance of two prominent LLMs on two benchmark datasets; (3) We compare our method against both supervised fine-tuned and zero-shot LLMs, as well as three established baseline models (Flan-T5, BART, and GPT4-o). To gain deeper insights into our results, we also present an analysis on the important factors of using the proposed algorithm in practice.

2 Related Work

The research work in this paper is related to prior work at least in the following two dimensions: aspect-based text summarization and contrastive learning.

Aspect-based summarization. As previously described, the motivation behind aspect-based summarization is to extract and summarize information relevant to a given aspect from a document. For instance, one early work by Paul et al. (2010) aimed to summarize contrasting opinions to enhance the comprehensiveness of the final summary.

Recent research in aspect-based summarization can be broadly categorized into two areas: developing novel learning strategies and constructing new datasets. Regarding learning strategies, Gunel et al. (2023b) proposed a four-step approach, encompassing aspect extraction and merging, with each step employing a specialized model, such as a fine-tuned T5 model for aspect extraction. Ding et al. (2024) augment continual learning for aspect-based sentiment analysis. Tang et al. (2024) use in context learning and aspect-based sentiment analysis for review summarization through the generation of Key Points. Addressing challenges like missing aspects, Li et al. (2023) introduced an unsupervised method for extracting opinions from source documents for summarization. Unlike these previous studies, this paper focuses on improving the summary accuracy of aspect-based summarization. Specifically, for a given aspect, the proposed approach ensures that the generated summary is relevant to that particular aspect and not to other aspects. In addition, recent research has shown that LLMs cannot handle aspect-based summarization as well as traditional summarization tasks (Yang et al., 2023a), which is echoed in our baseline experiment (section 5) and further motivated to develop new learning strategies.

Concurrent with the development of new learning strategies, significant effort has been dedicated to creating benchmark datasets for aspect-based summarization. These datasets span various domains, including scientific publications (Takeshita et al., 2024), climate change (Ghinassi et al., 2024), social media (Zhan et al., 2022a), legal decisions (T.y.s.s. et al., 2024), disordered texts (Guo and Vosoughi, 2024), and news articles (Ahuja et al., 2022; Tan et al., 2020b). While many datasets are limited in the number of aspects or specific domains to focus on, some works have explored a larger range of aspects (Tan et al., 2020b) or not targeted on specific domains (Yang et al., 2023b). In this work, we utilize the ANYASPECT dataset (Tan et al., 2020b) and the COVIDET dataset (Zhan et al., 2022a) to ensure evaluation across diverse domains and document lengths.

Contrastive learning for summarization. While contrastive learning has been surveyed generally within the domain of natural language processing (Zhang et al., 2022) and extensively employed to enhance representation learning in text summarization (Xu et al., 2022), its application to aspect-based summarization remains largely unexplored. Specifically, contrastive learning has been shown to improve the alignment of generated summaries with source documents, reducing factual inconsistencies and hallucinations (Cao and Wang, 2021; Liu et al., 2022). It has also been used to address the issue of exposure bias (Sun and Li, 2021).

Zheng et al. (2021) further explored different strategies for constructing contrastive examples, such as masking, swapping, and replacing words or sentences, to improve learning performance in text summarization. Zhuang et al. (2024) proposed an automated method of constructing "hard" negative examples for contrastive learning. This method is applied within standard text summarization–our method similarly creates "hard" negative examples by leveraging negative aspects and summaries from the same source document as the positive example, thus reducing potential noise.

Liu and Liu (2021) proposed a novel approach that leverages contrastive learning to formulate summarization generation as a reference-free evaluation problem, where the model is trained to distinguish between high-quality and low-quality summaries without relying on reference summaries. Liu et al. (2024) utilize LLMs as evaluators in contrastive learning for smaller models such as BART. Wu et al. (2020) create a new evaluation method for text summarization without using gold standard summaries. They utilize linguistic and semantic aspects to perturb text summaries for negative samples and train their evaluator using contrastive learning. Chern et al. (2023) mitigate hallucinations and non-factual information in text summaries by implementing a contrastive reward learning framework that generates candidate summaries from a pretrained sequence-to-sequence model, which are then ranked using factuality metrics. Feng et al. (2024) utilize Contrastive Preference Optimization to mitigate hallucinations made by LLMs in news summarization.

Among existing works, Wang and Wan (2021) presents the most closely related study, employing contrastive learning to inform the (dis)similarity of aspects. However, the present work differs significantly in both loss function design and training strategy. Specifically, the proposed method defines the contrastive loss directly on the generated aspect-based summaries, compelling large language models to capture the crucial aspect-related information from long input documents.

3 Contrastive Fine-tuning

In aspect-based text summarization, the length and content variability of documents create a difficult learning environment for models. A successful model must identify and extract information specifically related to the given aspect and synthesize a summary. If the model cannot effectively distinguish between informative sentences, it risks generating a generic summary that omits aspect-specific information, thus failing to fulfill its objective. To facilitate the model's learning of pertinent aspect-related information, we employ contrastive learning through a particular design on loss function and example construction.

As with other applications of contrastive learning in text summarization, two key components are essential: the definition of contrastive loss and the construction of contrastive examples. In this work, to mitigate hallucinations, we utilize only the reference summary of each example as the positive example. Consequently, our focus in example construction is solely on generating negative examples.

For each training example consisting of a document d, aspect a, and summary s, the formulation of contrastive learning for aspect-based text summarization can be expressed as:

$$\mathcal{L}_s = -\big[\log p(s \mid d, a) - \frac{\lambda}{n}\sum_{k=1}^{n}\log p(s_k^- \mid d, a)\big] \tag{1}$$

where s_k^- represents a contrastive summary given the same document d and aspect a, n is the number of negative examples per case, and λ is the contrastive coefficient that balances the original training loss and the contrastive loss. We recommend $\lambda \in (0, 1)$, and a more detailed discussion will be presented in section 6. As the model is trained to differentiate the correct summary from its contrastive example with the same aspect, we named this method as **aspect-anchored contrastive learning** with an example shown in Figure 2.

Unlike losses defined in previous work (Oord et al., 2018), the contrastive loss in Equation 1 can be interpreted as the geometric mean of the likelihoods of the negative examples, $\log(\prod_{k=1}^{n} p(s_k^- \mid$

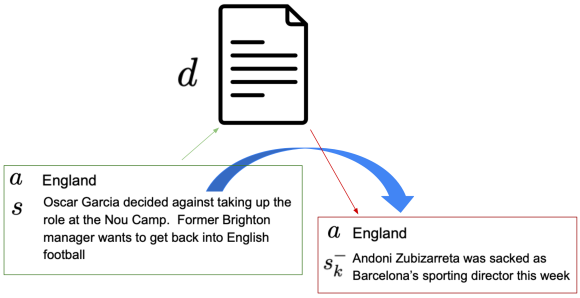


Figure 2: Illustration of *aspect-anchored* contrastive example generation from the AnyAspect dataset (Tan et al., 2020a). Given the document and the aspect, a contrastive example is constructed by selecting a summary from the same source document with a different aspect.

$d, a))^{1/n}$, which is less sensitive to extreme cases where the loss of a single negative example might dominate the others.

Variants of the Loss. Another variant of the loss function focuses on the aspect dimension rather than the summary. The corresponding loss function is defined as:

$$\mathcal{L}_a = -\Big[\log p(s \mid d, a) - \frac{\lambda}{n}\sum_{k=1}^{n} \log p(s \mid d, a_k^-)\Big] \tag{2}$$

where a_k^- is a random aspect from the the same document such that $a_k^- \neq a$. Unlike $\mathcal{L}_s$, $\mathcal{L}_a$ is defined directly on the aspect while maintaining the same summary. Intuitively, this loss function encourages the model to focus on the contrastive comparison between the original aspect and the negative aspects. However, in practice, we observed that this loss often confuses the model during generation, as the outputs for both the original loss and the contrastive loss are identical—both are s. Following the same naming convention, we name this method as **summary-anchored contrastive learning**.

A further variant could combine $\mathcal{L}_s$ and $\mathcal{L}_a$, applying negative examples to both summary and aspect variables. However, this poses additional training challenges, so we chose not to explore this direction further.

Quality of Negative Examples. In this work, for a given document and aspect, we select the summary from another aspect within the same document as the negative example in $\mathcal{L}_s$. Similarly, for $\mathcal{L}_a$, we choose the aspects available in the same document as the negative examples. By augmenting the summaries and aspects from within the same data point, we ensure that these negative examples are on-topic and semantically sound.

We noticed that prior work on contrastive learning often struggles with the quality and various issues when constructing negative examples, as discussed in Zhang et al. (2022). For example, noise introduced into the negative examples may cause further issues in a text summarization system, such as hallucinations (Ji et al., 2023). On the other hand, the proposed example construction method helps the model learn which information is similar and dissimilar from the same input document without the risk of noise from automatically perturbed samples.

4 Experimental Setup

This section describes the experimental setup, including datasets, baseline and competitive models, evaluation metrics, and additional implementation details.

4.1 Datasets

We run our experiments on two aspect-based text summarization datasets: CovidET (Zhan et al., 2022b), and AnyAspect (Tan et al., 2020a). We choose two datasets with vastly different domains to ensure that our method can perform well on datasets of varying content. Details about dataset size can be referenced at Table 4.

CovidET consists of sentiment aspect-based text summarization of 1,900 Reddit posts from r/COVID19_support. There are seven emotion-based aspects: ANGER, ANTICIPATION, JOY, TRUST, FEAR, SADNESS, and DISGUST. The Reddit posts range from 50 – 500 tokens, averaging at 100 tokens.

AnyAspect is derived from the popular CNN/DailyMail (Hermann et al., 2015) dataset, in which a named entity recognition model was used to extract aspects from the pre-existing data and formulate an aspect-based text summarization dataset. The original dataset size has over 2 million training examples with more than 339 thousand aspects, with each document averaging around 680 tokens. To make sure the evaluation focuses on contrastive learning instead of a large number of aspects, we selected a subset by taking the top ten most frequent aspects. In addition, we removed aspects that are synonymous to obtain a set of ten unique aspects. This yielded the following aspects

in our dataset partition: ENGLAND, U.S., OBAMA, CITY, PERSON, COUNTRY, CHINA, FACEBOOK, SCOTLAND, and SPAIN.

4.2 Baseline and Competitive Models

We evaluated the performance of several prominent large language models (LLMs) under three distinct training paradigms: contrastive learning, supervised fine-tuning, and zero-shot learning. Our selection of LLMs focused on widely used and representative open-source families with varying parameter sizes to ensure the generalizability of our findings. Specifically, we chose:

- Llama 2 (7B & 13B) (Touvron et al., 2023): Llama 2 is a family of open-source LLMs developed by Meta.
- Pythia (1B & 6.9B) (Biderman et al., 2023): Pythia is a suite of open-source language models trained by EleutherAI.

To establish robust performance baselines, we also evaluated three established models known for their strong performance in text generation tasks.

- Flan-T5 (Chung et al., 2022): Flan-T5 is a T5 model (Raffel et al., 2020) with instruction tuning that has demonstrated exceptional zero-shot performance across a wide range of text generation tasks.
- BART (Lewis et al., 2019): BART is another widely used sequence-to-sequence model particularly effective for text summarization.
- GPT-4o (OpenAI, 2024): GPT-4o is a decoder-only language model that achieves high performance on a wide range of language tasks.

By including these models, we establish a strong foundation for comparison, allowing us to effectively assess the improvements offered by our proposed contrastive learning approach when applied to LLMs.

For all the models used in the experiment, please refer to Appendix B for the model cards.

4.3 Evaluation metrics

We evaluate the quality of the summary through two categories of evaluation metrics: traditional methods in the form of ROUGE and BLEU scores as well as an LLM critique through Llama 3 (Grattafiori et al., 2024). We follow Mullick et al. (2024)'s method of using an LLM for evaluation by breaking down the critique prompt into four categories:

- Relevance (Rel): how well the content of the summary captures the aspect.
- Coverage (Cov): the extent to which the model-generated summary captures all of the pertinent information in the reference summary.
- Impurity (Imp): how well the model separates the aspects by not containing any information pertaining to a different aspect.
- Rating (Rat): a general rating of the quality of the summary with consideration of how clear, concise, accurate, and engaging the summary is.

We differ from their setup in that we use Llama 3 as our evaluator as opposed to GPT-4. In order to ensure our results were not influenced by an evaluator critiquing summaries output by the same model, we used a model that was not implemented in our experimental setup. We evaluate the highest performing contrastive fine-tuned and standard fine-tuned models compared to the baselines.

4.4 Implementation Details

All fine-tuning experiments were conducted using the LoRA (Low-Rank Adaptation) algorithm (Hu et al., 2021) with a rank $r = 8$ and scaling factor $\alpha = 16$. This parameterization was chosen based on preliminary experiments to balance performance and computational efficiency. We evaluated several model configurations: baseline LLMs (*zero-shot*), fine-tuned LLMs (*supervised fine-tuning*), and fine-tuned LLMs incorporating our proposed contrastive loss (*contrastive fine-tuning*).

We performed a hyperparameter search and used the following hyperparameters on the AnyAspect dataset: cutoff length as 500, λ as 0.25, learning rate as 10^{-4}. Following a similar procedure, we used the following hyperparameters on the CovidET dataset: cutoff length as 300, λ as 0.25, learning rate as 10^{-5}. We discuss the effects of changing these hyperparameters in detail in section 6. We fine-tuned our models on the AnyAspect subset for five epochs. This number of epochs was empirically determined by observing convergence behavior during initial experiments. For the smaller CovidET dataset, we fine-tuned for 10 epochs, also guided by empirical observations of convergence. All experiments were run on two NVIDIA A100 GPUs. This setup enabled us to train and evaluate our models within a reasonable time frame.

5 Experimental Results

Our experiments demonstrate the effectiveness of contrastive learning for aspect-based text summarization, as shown with the AnyAspect dataset in Table 1. Due to the page limit, the results of the CovidET dataset (in Table 5), along with other additional results, are included in Appendix C.

Across both the AnyAspect and CovidET datasets, our contrastive fine-tuning approach outperformed baseline models, standard fine-tuned LLMs, and zero-shot LLMs on most of the evaluation dimensions. Compared to standard fine-tuning, we notice that our method has a stronger performance on the AnyAspect dataset compared to CovidET. This indicates that contrastive fine-tuning may be a more viable approach depending on the data.

The performance improvement on AnyAspect is nearly 2% on ROUGE-1 score, 0.5% improvements on ROUGE-2, over 1% improvement in ROUGE-L, and over 2% improvement in BLEU score with the Llama 2 (13B) model when trained with our contrastive loss compared to its non-contrastive counterpart. We validate our findings with significance testing between Llama 2 (13B) contrastive fine-tuned and Llama 2 (13B) supervised fine-tuned for AnyAspect and find a p-value of 0.003. A table of p-values of the top-performing contrastive fine-tuned models can be referenced at Table 8. As shown in a specific example (in Table 2), the generated summary from the contrastive fine-tuned model is concise and directly references the specified aspect.

We also confirmed the significant advantage of fine-tuning over zero-shot inference, with fine-tuned models exhibiting substantial gains. While encoder-decoder models like Flan-T5 and BART provided strong baselines, our contrastive fine-tuned LLMs achieved superior performance. Finally, we observed a performance difference across datasets, which we attribute to variations in domain and aspect granularity.

In the remainder of this section, we will dive into different perspectives of the results. Most of the discussion will be based on ROUGE and BLEU, while the results from LLM-based evaluation will be discussed in the end of the section. An analysis of negative example construction will be presented in section 6 with other hyper-parameters.

Performance across Different Models. Among the contrastively fine-tuned LLMs, Llama 2 consistently outperformed Pythia, reinforcing previous findings on LLM performance hierarchies in aspect-based summarization (Mullick et al., 2024). This superior performance of Llama 2 may be attributed to its larger context size and extensive pre-training data. Furthermore, the substantial performance gap between fine-tuned and zero-shot LLMs (over 10% in ROUGE-1) confirms the benefits of data-specific adaptation.

Regarding model size, we see that Llama 2 13B outperforms its smaller counterpart for the AnyAspect dataset. The same is not observed for Pythia, where Pythia 1B outperforms its larger counterpart. For CovidET, we see Llama 2 7B and 13B's results are commensurate, whereas the larger Pythia model outperforms the smaller.

While Flan-T5, BART, and GPT-4o provided competitive baselines, their performance trailed that of the contrastively fine-tuned Llama 2, suggesting that the combination of LLM scale and contrastive learning provides the best performance. The similar performance of Flan-T5 and BART can be attributed to their shared encoder-decoder architecture, commonly recognized for its effectiveness in text summarization tasks. Further investigation is warranted to explore the relative strengths of

	R1	R2	RL	BLEU
Baselines				
BART	26.3	10.8	19.2	21.2
Flan-T5	26.8	9.9	19.6	23.1
GPT-4o	17.9	5.9	11.9	11.0
Zero-shot				
Pythia (1B)	14.6	3.5	10.1	7.6
Pythia (6.9B)	14.6	3.6	10.1	10.4
Llama 2 (7B)	13.1	3.7	9.5	11.6
Llama 2 (13B)	15.6	5.1	11.0	14.2
Supervised FT				
Pythia (1B)	24.1	7.9	17.9	25.9
Pythia (6.9B)	25.7	8.9	19.3	27.8
Llama 2 (7B)	29.0	10.7	21.6	30.0
Llama 2 (13B)	30.3	11.7	22.6	30.5
Proposed Method: Contrastive FT - Summary anchored				
Pythia (1B)	20.9	6.3	15.7	17.9
Pythia (6.9B)	19.6	5.5	14.6	18.0
Llama 2 (7B)	25.4	8.7	18.8	22.8
Llama 2 (13B)	24.7	8.5	18.9	22.9
Proposed Method: Contrastive FT - Aspect anchored				
Pythia (1B)	26.2	8.9	19.4	26.9
Pythia (6.9B)	25.8	8.9	19.6	25.3
Llama 2 (7B)	30.5	12.1	22.7	31.6
Llama 2 (13B)	**31.6**	**12.2**	**23.9**	**32.7**

Table 1: The evaluation results on the AnyAspect dataset. The labels "Aspect anchored" and "Summary anchored" represent the aspect-anchored and summary-anchored contrastive learning, respectively.

decoder-only versus encoder-decoder architectures for text summarization.

Performance Difference between Datasets. We observed a performance difference between the datasets, with models generally achieving higher scores on AnyAspect compared to CovidET. This discrepancy likely stems from domain differences and aspect granularity. AnyAspect, derived from CNN/Daily Mail news articles, aligns more closely with the training data of models like BART (which was instruction-tuned on CNN/Daily Mail). CovidET, focusing on sentiment analysis, presents a different domain. Additionally, the nature of the aspects themselves may play a role. The aspects from the AnyAspect dataset (e.g., SPAIN vs. U.S.) are more distinct than those in CovidET (e.g., JOY vs. ANTICIPATION), potentially making the latter a more challenging classification task. More nuanced aspects could increase the difficulty of creating distinct summaries, leading to lower evaluation scores.

LLM-based Evaluation Metrics. Unlike ROUGE and BLEU that are more consistent across different dimensions, LLM-based Evaluation (as in Table 7), on the other hand, revealed different patterns. Contrary to the results obtained with ROUGE and BLEU, LLM-based metrics tended to assign higher scores to summaries generated for the CovidET dataset than those for AnyAspect. This suggests that while the summaries produced for AnyAspect might be closer to the reference summaries in terms of n-gram overlap (as reflected in ROUGE scores) the summaries generated for CovidET, though potentially differing in wording, may excel in facets other than word-for-word similarity. This underscores the need to consider multiple evaluation perspectives to fully understand the strengths and weaknesses of different summarization models.

Human Evaluation. We conducted a human evaluation in order to verify which of the two highest performing methods (standard fine tuning or contrastive) was able to create the highest quality summaries. We surveyed six computer science graduate students all of whom are native English speakers. We presented the evaluators with ten sample data points from AnyAspect along with two summaries of the article – one generated from the supervised fine-tuned Llama 2 7B model and the other from our contrastive fine-tuned Llama 2 7B model with an aspect anchor. We asked them to choose which summary was the best in regards to quality and coverage of the aspect. An example question from the survey can be referenced in Figure 5. The evaluators selected our contrastive fine-tuned summaries as superior for six summaries, the standard fine-tuned for two summaries, and tied for two other questions. Our findings conclude that human evaluators on whole found the contrastive fine-tuned summaries to be of higher quality, solidifying our findings.

6 Algorithm Analysis

To gain a comprehensive understanding of the proposed algorithms, we used Llama 2 7B as the base model and AnyAspect as the dataset. We considered the following hyper-parameters and their effect on performance: learning rate, strength of the negative example via contrastive coefficient, and cutoff length (as shown in Figure 3). Based on the results in Table 1 and Table 5, we also analyze the impact of negative example construction.

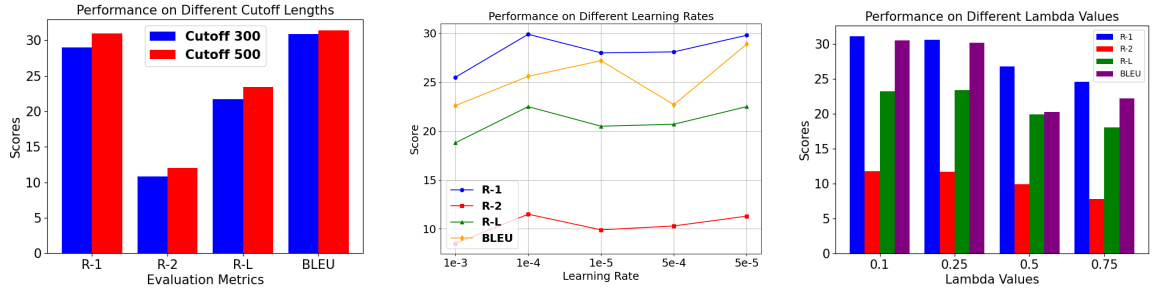
Learning Rate. The following learning rates were evaluated: 1e-3, 1e-4, 1e-5, 5e-4, and 5e-5. We can see that Llama 2 peaks in performance at 1e-4 with diminishing returns for both large and smaller learning rates. This indicates that 1e-4 strikes a balance between performance, convergence, and stability when performing contrastive fine-tuning.

Contrastive Coefficient λ. We evaluate the following values of λ in our contrastive loss ranging between (0,1): 0.1, 0.25, 0.5, and 0.75. Our results indicate a linear decrease in ROUGE and BLEU scores as λ value increases. The smallest λ value (0.1) yielded the highest ROUGE-1 score among all evaluated hyper-parameters in our algorithm analysis. This indicates that a positive example supplemented by a weaker impact from negative examples was more helpful in the model's learning as opposed to more dominant negative examples during training.

Cutoff Length. We evaluate two different cutoff lengths during training and inference: 300 tokens vs. 500 tokens – the upper limit is capped by the hardware resource we used in the experiment. We see performance increase as cutoff length increases. Due to the longer length of the input document for a news domain, some content related to the aspects will be located closer to the end of the document. A

Instruction	*Generate a summary based on Aspect:* FACEBOOK
Zero-shot summary	The community of body builders are in shock after the premature death of celebrated competitor Phil Primmer.
Supervised FT summary	The body-building community is in shock after the premature death of celebrated competitor Phil Primmer. The Darwin man, aged just 42, had reportedly just been to see a doctor to seek treatment for neck pain when he died. He was given CPR but it was too late by the time emergency services arrived, reportedly to the gym he owned according to friends on Facebook.
Contrastive FT summary	His Facebook page has been inundated with heartfelt tributes from people inspired by 'an Australian body building legend who will never be forgotten.

Table 2: The generated summaries are based on an example from the AnyAspect dataset with the FACEBOOK. The source document is ignored due to the page limit. Even without the source document, it is clear that the generated summary from contrastive fine-tuned Llama 2 focuses more directly on the aspect without much irrelevant information.



(a) cutoff length (b) initial learning rate (c) λ

Figure 3: The performance under different hyper-parameters.

larger cutoff length will allow the model to encapsulate as much information as possible to create the summary, resulting in better model performance.

Negative Example Construction. We notice that the aspect-anchored contrastive fine-tuned LLMs performed better on the AnyAspect dataset, whereas the summary-anchored fine-tuned models fared better on CovidET. We attribute this to the difference in domain and granularity of distinction between the sets of aspects for each dataset. The aspects in CovidET were not as distinct, where the difference between each emotion may not be wholly obvious to the model. This explains why the models that were fine-tuned with a contrastive aspect fared better on CovidET, where the aspect set may have been a source of confusion. The summary-anchored CovidET dataset contained slightly less training examples (7,650) compared to the aspect-based (8,270), as shown in Table 4, yet the summary-anchored models still saw higher ROUGE and BLEU scores, further illustrating the value of the contrastive aspects in this particular dataset.

For AnyAspect, the model may have been able to encapsulate these differences without the need for training with contrastive aspects to be able to distinguish between the set. Within the news domain, there are a multitude of details encompassing each story, such as key figures, events, places, and time, which can lead to a variety of different summaries. This could explain why contrastive summaries were more helpful for the AnyAspect dataset.

7 Conclusion

In this study, we perform the first systematic analysis of LLM performance in aspect-based text summarization with a formulated contrastive loss function. Our method sees an improvement in evaluation scores compared to standard fine-tuning and baseline models. We contribute additional analyses in an ablation study which evaluates the effects of negative example construction on performance, as well as hyper-parameters such as cutoff length, learning rate, and strength of the contrastive component through a coefficient value.

8 Limitations

This work primarily explored the impact of negative example construction within a contrastive learning framework for aspect-based summarization. Future research could investigate the influence of positive example selection and generation on overall performance. Furthermore, while we demonstrated the effectiveness of our approach across two benchmark datasets and two large language models, a more extensive evaluation involving a wider range of datasets and model architectures would further solidify these findings.

Acknowledgments

The authors thank the anonymous reviewers for their insightful feedback and suggestions. This research was partially supported by the NSF award #2007492 and the UVA TYDE Seed grant awarded.

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2021. Aspectnews: Aspect-oriented summarization of news documents. *arXiv preprint arXiv:2110.08296*.

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-Oriented Summarization of News Documents. ArXiv:2110.08296 [cs].

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive Learning for Improving Faithfulness and Factuality in Abstractive Summarization. ArXiv:2109.09209 [cs].

I-chun Chern, Zhiruo Wang, Sanjan Das, Bhavuk Sharma, Pengfei Liu, and Graham Neubig. 2023. Improving factuality of abstractive summarization via contrastive reward learning. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 55–60. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Xuanwen Ding, Jie Zhou, Liang Dou, Qin Chen, Yuanbin Wu, Arlene Chen, and Liang He. 2024. Boosting Large Language Models with Continual Learning for Aspect-based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4367–4377, Miami, Florida, USA. Association for Computational Linguistics.

Huawen Feng, Yan Fan, Xiong Liu, Ting-En Lin, Zekun Yao, Yuchuan Wu, Fei Huang, Yongbin Li, and Qianli Ma. 2024. Improving Factual Consistency of News Summarization by Contrastive Preference Optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11084–11100, Miami, Florida, USA. Association for Computational Linguistics.

Iacopo Ghinassi, Leonardo Catalano, and Tommaso Colella. 2024. Efficient Aspect-Based Summarization of Climate Change Reports with Small Language Models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 123–139, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Beliz Gunel, Sandeep Tata, and Marc Najork. 2023a. Strum: Extractive aspect-based contrastive summarization. In *Companion Proceedings of the ACM Web Conference 2023*, pages 28–31.

Beliz Gunel, Sandeep Tata, and Marc Najork. 2023b. STRUM: Extractive Aspect-Based Contrastive Summarization. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pages 28–31, New York, NY, USA. Association for Computing Machinery.

Xiaobo Guo and Soroush Vosoughi. 2024. Disordered-DABS: A Benchmark for Dynamic Aspect-Based Summarization in Disordered Texts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 416–431, Miami, Florida, USA. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. Aspect-aware Unsupervised Extractive Opinion Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12662–12678, Toronto, Canada. Association for Computational Linguistics.

Wei Liu, Huanqin Wu, Wenjing Mu, Zhen Li, Tao Chen, and Dan Nie. 2022. CO2Sum:Contrastive Learning for Factual-Consistent Abstractive Summarization. ArXiv:2112.01147 [cs].

Yixin Liu and Pengfei Liu. 2021. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. ArXiv:2106.01890 [cs].

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8647–8664. Association for Computational Linguistics.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. Leveraging the power of LLMs: A fine-tuning approach for high-quality aspect-based summarization. Version Number: 1.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2024. Gpt-4o technical report. Accessed: 2025-05-13.

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing Contrastive Viewpoints in Opinionated Text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, MA. Association for Computational Linguistics.

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.

Shichao Sun and Wenjie Li. 2021. Alleviating Exposure Bias via Contrastive Learning for Abstractive Text Summarization. ArXiv:2108.11846 [cs].

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024. ACLSum: A New Dataset for Aspect-based Summarization of Scientific Publications. ArXiv:2403.05303 [cs].

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020a. 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. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020b. 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, Online. Association for Computational Linguistics.

An Tang, Xiuzhen Zhang, Minh Dinh, and Erik Cambria. 2024. Prompted Aspect Key Point Analysis for Quantitative Review Summarization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10691–10708, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Santosh T.y.s.s., Mahmoud Aly, and Matthias Grabmair. 2024. LexAbSumm: Aspect-based Summarization of Legal Decisions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10422–10431, Torino, Italia. ELRA and ICCL.

Ke Wang and Xiaojun Wan. 2021. TransSum: Translating Aspect and Sentiment Embeddings for Self-Supervised Opinion Summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 729–742, Online. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3612–3621, Online. Association for Computational Linguistics.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence Level Contrastive Learning for Text Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11556–11565. Number: 10.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023a. Exploring the limits of ChatGPT for query or aspect-based text summarization. Version Number: 1.

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.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022a. Why Do You Feel This Way? Summarizing Triggers of Emotions in Social Media Posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022b. Why do you feel this way? summarizing triggers of emotions in social media posts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9436–9453. Association for Computational Linguistics.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, Ling Fan, and Zhe Wang. 2021. Enhanced Seq2Seq Autoencoder via Contrastive Learning for Abstractive Text Summarization. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1764–1771.

Haojie Zhuang, Wei Emma Zhang, Chang Dong, Jian Yang, and Quan Sheng. 2024. Trainable hard negative examples in contrastive learning for unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1589–1600, St. Julian's, Malta. Association for Computational Linguistics.

A Prompt

The prompt template used to generated aspect-based summary in Python format.

```
f"Generate a summary based on

### Aspect: {aspect}

### Input: {input}

### Output:"
```

B Model Cards and Data Statistics

The sources of the models used in this work are listed in Table 3. The basic information of the datasets is presented in Table 4.

C Additional Results

We present the following additional results. Table 5 contains the results from our main evaluation of the four training paradigms on the CovidET dataset. Table 6 displays the results from our empirical study on the effects of various hyper-parameters. Table 7 contains the results from our LLM evaluation of the standard fine-tuned and contrastive fine-tuned results compared to baselines with Llama 3. Table 8 contain the p-values of some of the top performing contrastive models compared to their standard fine-tuned counterparts.

Model	Size	Link
Llama 2	7B	https://huggingface.co/meta-llama/Llama-2-7b-hf
Pythia	6.9B	https://huggingface.co/EleutherAI/pythia-6.9b
BART	406M	https://huggingface.co/facebook/bart-large-cnn
Flan-T5	780M	https://huggingface.co/google/flan-t5-large

Table 3: The model card links to all the models used in the experiments.

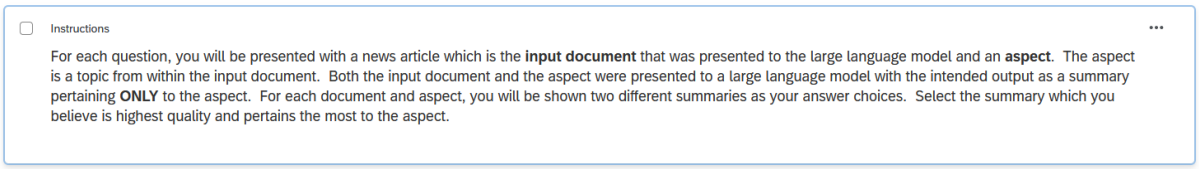


Figure 4: The instructions presented to the participants.

Figure 5: An example question from human analysis of text summary quality.

Partition	CovidET	AnyAspect
Original Training Set	4,188	25K
Aspect Anchored Training Set	8,270	50K
Summary Anchored Training Set	7,650	50K
Validation Set	1,524	2,122

Table 4: Statistics of the CovidET and AnyAspect datasets for training and evaluation.

	R1	R2	RL	BLEU
Baselines				
BART	18.9	3.2	12.6	11.8
Flan-T5	13.3	1.7	10.2	10.2
GPT-4o	10.9	1.7	7.7	6.8
Zero-shot				
Pythia (1B)	12.1	1.7	8.8	7.6
Pythia (6.9B)	11.5	1.3	8.2	7.4
Llama 2 (7B)	12.4	1.8	9.0	8.5
Llama 2 (13B)	10.0	1.4	7.4	8.3
Supervised fine-tuning				
Pythia (1B)	22.8	5.1	18.3	19.4
Pythia (6.9B)	24.2	5.9	19.3	21.1
Llama 2 (7B)	25.3	6.4	19.6	21.5
Llama 2 (13B)	25.3	6.4	20.0	22.3
Proposed Method: Contrastive fine-tuning - Summary anchored				
Pythia (1B)	22.6	5.0	17.9	18.9
Pythia (6.9B)	23.8	5.5	18.7	20.2
Llama 2 (7B)	25.6	6.4	20.0	21.8
Llama 2 (13B)	25.6	6.5	20.0	22.8
Proposed Method: Contrastive fine-tuning - Aspect anchored				
Pythia (1B)	22.5	5.1	18.0	19.0
Pythia (6.9B)	23.8	5.5	18.6	20.1
Llama 2 (7B)	25.1	6.2	19.7	21.8
Llama 2 (13B)	25.2	6.1	19.6	22.4

Table 5: The evaluation results on the CovidET dataset. The labels "Aspect anchored" and "Summary anchored" represent the aspect-anchored and summary-anchored contrastive learning, respectively.

Hyper-parameter	Value	R-1	R-2	R-L	BLEU
Cutoff Length	300	29.0	10.8	21.7	30.9
	500	30.9	12.0	23.4	31.1
Learning Rate	1e-3	25.5	8.5	18.8	22.6
	1e-4	29.9	11.5	22.5	25.6
	1e-5	28.1	9.9	20.5	27.2
	5e-4	28.1	10.3	20.7	22.7
	5e-5	29.8	11.3	22.1	28.9
Lambda	0.1	31.1	11.8	23.2	30.5
	0.25	30.6	11.7	23.4	30.2
	0.5	26.8	9.9	19.9	20.3
	0.75	24.6	7.8	18.1	22.2

Table 6: Performance for different hyper-parameters on AnyAspect aspect contrastive anchor with Llama 2 (7B).

Partition	**Model**	**Rel**	**Cov**	**Imp**	**Rat**
AnyAspect	Flan T-5 (baseline)	45.9	30.7	94.9	45.6
	Llama 2 13B (Standard Fine-Tuned)	53.8	35.5	95.4	40.3
	Llama 2 13B (Contrastive - Aspect)	52.4	28.3	95.6	36.0
CovidET	BART (baseline)	59.4	30.7	74.2	52.9
	Llama 2 13B (Standard Fine-Tuned)	79.9	36.4	62.0	50.9
	Llama 2 13B (Contrastive - Summary)	82.1	35.6	61.2	51.8

Table 7: Llama 3 Evaluation of Top Performing Models.

Dataset	Model	P-value
AnyAspect	Llama 2 (13B)	0.003
	Llama 2 (7B)	0.000008
	Pythia (1B)	0.000001
CovidET	Llama 2 (13B)	0.395
	Llama 2 (7B)	0.377

Table 8: P-values of top performing contrastive fine-tuned models compared to their standard fine-tuned counterparts