

From Information to Insight: Leveraging LLMs for Open Aspect-Based Educational Summarization

Anonymous ACL submission

Abstract

This paper addresses the challenge of aspect-based summarization in education by introducing Reflective ASpect-based summarization (ReflectASP), a dataset that summarizes student reflections on STEM lectures. Despite the promising performance of large language models in general summarization, their application to nuanced, aspect-specific summaries in educational texts remains under-explored. ReflectASP eases the exploration of open-aspect-based summarization (OABS), overcoming the limitations of current datasets and annotation complexities. We leverage GPT-4 for generating reference summaries and propose a self-refine framework to enhance summary quality. Our work benchmarks the capabilities of different language models in this novel context, contributing a unique dataset and insights into effective summarization strategies for educational content. We will make our model, dataset, and all human evaluation results available at urlannonymized_for_review.

1 Introduction

Real-world documents often contain various *aspects* (Titov and McDonald, 2008), necessitating summaries that respond to specific user interests. While aspect-based summarization (ABS) covers shareable subtopics for different documents, such as *Early Life and Career* in Wikipedia biographies (Hayashi et al., 2021) and *Penalty* in fraud news articles (Ahuja et al., 2022), the emergence of open-aspect-based summarization (OABS) (Tan et al., 2020; Yang et al., 2023; Amar et al., 2023) allows for identifying unique aspects for each document and generating summaries accordingly.

However, the development of OABS faces challenges, particularly due to the limited quality of existing datasets (mainly in news and Wikipedia articles) and the complexities in collecting manual annotations for aspect-based summaries. For instance, OPENASP (Amar et al., 2023) suffered

All reflections (23 students)

- The initial set up. Specifically the limits of integration and the variable of integration.
- I felt the lecture was going too fast and felt that we should go over details more rather than flying through each tophat question.
- How to determine the setup for the charges.
- I am somewhat confused on what dq/ds represent in the problems in parts 3/4 of this lecture series.
- I am really having trouble setting up integrals for the electric field. I don't understand how to set it up given the linear density and the charge.
- The limits on the integrals confused me the most. I did not know why some of them changed midway through the problem.
- Ring of charge and the symmetry cancellation.
- the integrations were confusing but i think i just need to focus more and review calc
- I am confused on how to figure out what dq should be in example problems that use that equation
- Determining when to use cos or sin for integration of a ring of charge. Also drawing $d\vec{E}$ vectors along the z-axis was a bit confusing for me.
- The class problem regarding finding the x component of Electric field.
- While I understand the solving the integral process, I could use more clarity/examples of setting up integrals in different situations. Also, a deeper explanation of what exactly is happening to the electric field when influenced by an extended charge.
- ... (Omitted the remaining)

Human Generated Summary

Prompt: Given students' responses, create a short summary using your own words (~40 words) of it. The summary needs to be a coherent paragraph and should include the major points.

Many of the students had conceptual trouble today. Most of them mentioned trouble with integration and problem set up, while others mentioned the actual concepts such as ring of charge, electric fields, linear charge, etc. Some students had issues with the examples they were doing in class as well.

Aspect-Based: Integration

Prompt: Please write a short summary with no more than 100 words, focusing on the topic of Integration based on below reflections

LLM Generated:

Many he students found the integrations to be confusing, particularly the initial set up and the limits of integration. They also struggled with determining the variable of integration and the symmetry cancellation. Some students felt that the lecture was going too fast and that more examples were needed to fully understand the concepts. The students were also confused about the example problems, specifically the ring of charge calculations and the x component of Electric field. They had trouble setting up integrals for the electric field and understanding the sanity checks. The dipole questions were also challenging for some students. Overall, the students felt that more examples and clarity were needed to fully understand the topics covered in the lecture.

Figure 1: An example aspect-based summary of students' confusing points about integration in a college-level Physics course. While the human-generated summary focuses on summarizing all reflections and lacks details on specific aspects, our model generates an aspect-focused summary with reasons/details, which is also **bolded** in the original reflections.

methodological challenges while collecting manual annotations on aspects and guiding annotators to extract specific sentences from generic summaries for aspect-based summaries. These challenges arise in formulating reference summaries, making it difficult to evaluate the generation quality of models in an aspect-focused context.

This paper argues for domain-specific aspect construction and appropriate evaluations, focusing on opinions in the educational domain in the form of student reflections. Student reflections provide valuable insights into students' learning (Menekse et al., 2011) and help instructors identify student misconceptions (Aslan et al., 2019; Alrajhi et al., 2021; Jacobs et al., 2022), thereby enabling them to strategize suitable follow-up actions.

In the example in Figure 1, twenty-three students wrote reflections after a physics lecture. Looking at one aspect, “Integration”, the generic summary of major points (see human-generated summary) might be inadequate for an instructor interested in pinpointing particular confusions on integration or pairing a student eager to learn from appropriate peers. Instead, a more appropriate summary would be aspect-based, directly addressing the question “What do students find most confusing about ‘Integration’ in this lecture?”.

Student reflections encompass diverse aspects, and this domain is good for OASBS. Firstly, the concept of **aspects** in our context is inherently open-ended, varying significantly across subjects. For instance, discussions on “sorting algorithms” are majorly relevant to Computer Science courses, illustrating the specific variation across subjects. This diversity in aspects allows testing the model’s capability to capture the aspect-related information and accurately generate a summary. Furthermore, students articulate their thoughts on the same aspect in different details and points, thus demanding the model to identify the common obstacles. Third, exploring the capabilities of smaller open-sourced models for generating aspect-based summaries is crucial, especially considering resource constraints and the potential for future updates.

We introduce Reflective ASPECT-based summarization (ReflectASP), a novel dataset using REFLECTSUMM (Anonymous, 2023) to create. This dataset contains 1,064 data instances, including aspect, source reflections, annotated aspect-based clusters of student reflections, and silver GPT-4 generated summaries. We cast our task as an *open-aspect-based multi-document summarization*. The dataset comes with human-generated phrases to summarize the lecture, which we take as the open aspects, alongside instance-level labels for individual reflections to form clusters for each annotated aspect. Due to the lack of reference summaries, we utilized GPT-4 to generate silver reference summaries with aspect-based prompts to benchmark several open- and closed-source LLM’s performance in a reference-based setting. We further introduced a self-refine framework, which leveraged the LLMs’s capability to self-critic and improve (Madaan et al., 2023; Huang et al., 2023; Welleck et al., 2023), to enhance the quality of summaries. Furthermore, we conduct both human evaluation and automatic analysis to examine LLM’s capabil-

ity in generating aspect-based summarizations.

To sum up, our contributions are threefold:

1. We contribute a new dataset for open-aspect-based summarization, consisting of aspects and filtered source texts, together with GPT-4 generated silver references.
2. We use this dataset to benchmark LLM’s capability on aspect-based summarization, then introduce a self-refine framework to boost the performance.
3. We conduct human evaluations and perform a series of data-driven studies to answer the question: *what common strategies are used by large language models (LLMs) to improve the writing of aspect-based summaries*.

2 Related Work

Aspect-based summarization is a specialized form of text generation that produces summaries focused on specific subtopics or aspects of interest. Unlike traditional methods that often rely on extractive techniques, recent efforts in aspect-based summarization aim to generate abstractive summaries and have been conducted across different domains, including news (Frermann and Klementiev, 2019; Bahrainian et al., 2022) and Wikipedia (Hayashi et al., 2021). However, these approaches typically work on a limited set of pre-defined aspects: i.e., Ahuja et al. (2022) manually defined two aspects for each domain based on keyword searches. Restricted aspects make it challenging to apply such approaches to domains with more variable content, such as students’ reflections, where aspects can differ dramatically even within a single course. A few exceptions looked into open aspect-based summarization (OABS) (Tan et al., 2020; Amar et al., 2023), where aspects are document-based. Meanwhile, LLMs have demonstrated zero-shot capability in performing OABS tasks (Amar et al., 2023) in a multi-document setting. *Our study explores the use of LLMs for OABS, specifically focusing on the diverse, document-dependent aspects found in student reflections and assessing LLM performance through extensive experiments.*

LLM Feedback and Refinement It has been recognized that feedback at inference-time is essential for LLMs to refine their answers (Madaan et al., 2023; Welleck et al., 2023; Zheng et al., 2023; Huang et al., 2023; Gao et al., 2023; Pan et al.,

2023; Shinn et al., 2023). In the domain of text summarization, prior works studied training summarization models through human feedback in the form of numerical quality scores (Stiennon et al., 2020; Wu et al., 2021; Nguyen et al., 2022). Additionally, Liu et al. (2023a) proposed automatically correcting factual inconsistencies in generated summaries using generated feedback simulated to be from humans. Notably, one prior work leverages GPT-3.5 to iteratively revise summaries to improve the factuality and controllability in news articles (Zhang et al., 2023). We employ open-sourced LLMs to generate feedback based on minimal instructions, to produce more focused summaries on a given aspect. We further investigate the model’s generated feedback and examine the source of improvements from this refinement process.

3 Dataset

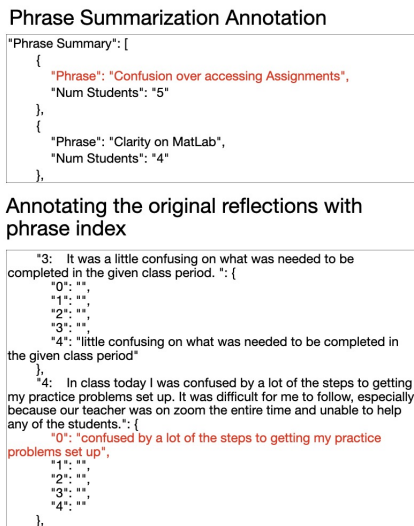


Figure 2: An example of phrase support annotation. Annotators are asked to produce at most five phrase summaries; they are then instructed to revisit the original reflections and assign a phrase index for each (In this example, the fourth reflection is annotated to be associated with the first phrase).

Dataset Curation The student reflections in REFLECTASP were exploited from REFLECTSUMM (Anonymous, 2023), which comes with reflections manually clustered by topic and general abstractive summaries. For each lecture, the dataset provides a collection of student reflections focusing on interesting or confusing points. One salient part of the dataset is the annotation on human-written phrase support, as shown in Figure 2. Annotators

are directed to mark original student reflections as evidence for their annotated noun phrases.

There were initially 782 reflection-summaries pairs in the dataset. For each lecture, annotators are instructed to provide five phrases summarizing the lecture reflection and how many students semantically mentioned each phrase. We construct our dataset for OASB by treating all reflections as the multi-document and the annotated phrases as the aspects. We removed lectures where the number of students was small (fewer than ten students, so summarization isn’t needed) and selected aspect-reflection pairs where at least five students mentioned the phrase. This reduced the total amount of data points from 3908 to 1096.

Aspect Analysis Out of the 1096 phrases, 778 are unique. To examine the variations among aspects, we encoded them using Phrase-BERT (Wang et al., 2021), followed by the application of the K-means unsupervised clustering algorithm, to organize them into clusters¹. Our analysis reveals several distinct groups of phrases. The primary group consists of course-specific terminologies, which vary across different courses and are dependent on the lecture and subject matter (i.e., Newton’s Laws in a Physics course). There are also multiple clusters of phrases that are shareable across different lectures, such as “Assignment related problems”, “Quiz and examination”, along with “Other Statements” and “No Confusions”.

The variability of aspects in the first group necessitates open aspects in aspect-based summarization to satisfy the user’s need to learn about interesting/confusing points. Moreover, we observe that reflections tagged with “No Confusion” carry the least amount of information and are deemed superficial. Thus, we excluded the data points with aspects annotated as “No confusion,” reducing the total number of data points to 1064. This refinement helps to focus on more substantive aspects.

Silver Reference Summary Constructing aspect-based reference summaries remains challenging. We employed OpenAI’s ChatGPT (GPT-4-turbo) as our LLM to execute zero-shot aspect-based summarization, similar to Zhang et al. (2023). For each case in the REFLECTASP dataset, we prompted the ChatGPT model to produce a focused summary centered around the aspect. (We include the prompt in Appendix B.1). The instructions emphasized

¹Details on the clustering process are in Appendix A.

	Domain	Collection	# of Instance	# Word/Input	# Doc/Input	# Aspect	Summ. Len.
FacetSum	Scientific	A	60,532	6,827	-	4	290
ASPECTNEWS	News	M	400	248	-	4	115
SPACE	Reviews	M	900	14,335	100	6	26
OASUM	Wikipedia	A	3,747,569	1,612	-	1,045,895	40
OpenASP	News	M	1,310	6,860	26	1,266	82
REFLECTASP (ours)	Reflections	M	1,064	480	30.2	1,064	87

Table 1: Descriptive statistics comparing prior similar datasets (top) to REFLECTASP. The first three are on ABS, and the others belong to OABS. For the **Collection** of aspects column, **A** stands for “Automatic” and **M** stands for “Manual”. **# Word/Input** (concatenated reflections/documents) means the average of the total number of words included in a summarization input. **# of Doc /Input** measures the average number of input reflections/documents/articles, and **Summarization length** is counted by words. The mark (-) refers to not applying to this dataset.

minimal requirements and explicitly requested the avoidance of unrelated text inclusions. These generated summaries serve as the silver reference for subsequent experiments. To evaluate the quality of the summaries, we include GPT-4 summaries as part of the human evaluations (Sec 5.2).

Compared to Other OABS Dataset Table 1 compares our REFLECTAPS to existing ABS and OABS datasets. Different from FacetSum (Meng et al., 2021) and OASUM (Yang et al., 2023), which relied on automatic methods for dataset curation, our dataset benefits from manual annotations by trained undergraduate students, retaining higher quality of data. Unlike OASUM and OpenASP, which might suffer from over-compression due to their dependency on sentences selected from generic summaries, similar to ASPECTNEWS, our dataset collected annotations of aspect-based clusters from input reflections, thus alleviating the risk of missing information from generic summaries.

4 Aspect-based Summarization Task

Given the entire set of course reflections from one lecture and an aspect such as “Integration”, we experiment with both pre-trained models and large language models and test how well they can pick up the salient reflections to generate an abstractive summary of the findings. When exposed to certain aspects, we aim to understand students’ more profound challenges or interests. We frame this task as an abstractive summarization task for two reasons: (1) Extractive summarization is straightforward, as the original reflections can be retrieved by matching the noun phrases. However, it may not fully capture the underlying themes of students’ reflections on specific aspects of their learning ex-

Role	Content
system:	You are a responsive abstractive summarizer that summarizes the collection of student lecture reflections by focusing on a specific topic.
user:	Please write a short summary with no more than 100 words, focusing on the topic of {topic} based on below reflections: {reflections} SUMMARY:

Table 2: The baseline prompt used for aspect-based summarization given an aspect.

perience. (2) The length of the summary, denoted as ‘S’, should be constrained within a reasonable range, particularly for lengthy inputs. Abstractive summarization accommodates the distillation of insights into a concise format.

4.1 Experimental Setup

For pre-trained models, we experimented with Longformer-Encoder-Decoder (LED) base model (Beltagy et al., 2020) for generating summaries. We additionally experiment with **Flan-T5-XL** (Chung et al., 2022), a strong baseline fine-tuned through instructions tuning on multiple tasks.

For LLM baselines, we include **LLAMA 2-7B chat** and **LLAMA 2-13B chat** (Touvron et al., 2023), and **Mistral-7B-Instruct-v0.1** (Mistral) (Jiang et al., 2023)² models for experiments. Similar to Amar et al. (2023), we use a fixed prompt to produce an aspect-based summary with the aspect being the annotated phrases (Table 2). We

²<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>

also include **GPT-3.5** as one strong baseline.³ All implementation details are in Appendix C.

4.2 Self-refine Method

In Figure 1, the middle sentences, such as “*Some students felt that the lecture was going too fast and that more examples were needed to fully understand the concepts.*”, are not related to the aspect of “Integration”. This indicated that LLMs might not fully understand instructions (Liu et al., 2023b) to concentrate on a specific aspect, resulting in generic summaries that encompass multiple aspects. Inspired by the success of recent lines of study on self-correction (Madaan et al., 2023; Welleck et al., 2023), we employ a *Generate-Suggest-Refine* framework to use the model to improve its outputs. More specifically, after **generating** an initial aspect-based summary, we prompt the model to **provide suggestions** to improve the summary by making it more concise and concentrated on the topic. We carefully craft the prompts to ensure the suggestions are grounded in the original reflections, whilst the revision suggestions should be based on the context of the first version. Lastly, we **refine** the summary by providing the LLM with all reflections, the initial draft, and improvement suggestions, prompting it to produce a **refined** version. Subsequent experimental results and analysis in later sections underscore the benefits and limitations of this proposed self-refine approach. We include the prompt in Appendix B.2. Our approach differed from prior work (Madaan et al., 2023; Welleck et al., 2023) in that they relied on few-shot samples and had restricted the feedback formatting. Instead, our work elicited the model’s capability to provide feedback and conducted extensive analysis to evaluate the quality of suggestions and refinement.

4.3 Evaluation Metrics

One challenge is the absence of gold-standard summaries for aspect-based tasks. We harnessed the annotated reflection clusters for the given aspect (details in Sec 3) and the GPT-4 generated silver summary. We posited that a good, relevant aspect-based summary should encapsulate contents covered in the annotated reflection cluster for a specific aspect. To quantify this, we measure the ROUGE F1 scores (Lin, 2004) (ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L)) between the system

outputs and the original reflections in the cluster. We further report the ROUGE F1 and BERTScore (Zhang* et al., 2020) between the GPT-4 generated reference and system generations.

5 Results

This section addresses two research questions: **RQ1**. Are LLMs capable of generating aspect-based summaries in a zero-shot setting? If so, how could we measure their quality? **RQ2**. How does the self-refine help with the summarization? We then conduct analyses based on the observations and perform a small-scale human evaluation.

5.1 Automatic Evaluation

RQ1. Table 3 shows that pre-trained models (LED and Flan-T5-XL) struggle to adhere to the instructions, resulting in lengthy summaries and lower quality when evaluated against GPT-4 summaries. When evaluated against annotated clusters, the higher ROUGE score can be attributed to the longer lengths of the generated summaries and their tendency towards near-extractive behavior in a zero-shot setting. Examples are in Appendix C.3. Upon examining LLM baselines (rows 3-6), LLAMA 2-7B obtains the highest ROUGE F1 score compared to cluster references among all four models. However, when evaluated against the GPT-4 silver summary, GPT-3.5 outperformed all three other models in ROUGE and BERTScore. This is unsurprising, given that they are both from the GPT family. LLAMA 2-13B excels in ROUGE F1 (0.6, 0.8, and 0.5 higher than Mistral concerning R-1, R-2, and R-L) than the other two open-sourced models. These differences are **significant**, with 95% confidence intervals. The variation in ROUGE scores (between the two references) can be attributed to the length of generated summaries, as the reference clusters can be as long as 500 words, thus favoring more extended system summaries. Finally, regarding BERTScore, GPT-3.5 outperforms all other baselines, and LLAMA 2-13B and Mistral significantly outperform LLAMA 2-7B.

Regarding **RQ2**, we assess the effects of our self-refine approach by comparing performances before and after refinement (row 3 vs. row 7, row 4 vs. row 8, row 5 vs. row 9, and row 6 vs. row 10) across all four LLM baselines. The performance of LLAMA 2 models dropped concerning all metrics except for LLAMA 2-13B BERTScore (i.e., the cluster ROUGE-L F1 dropped from 32.29 to

³GPT-3.5 turbo 1106 from <https://platform.openai.com/docs/models/gpt-3-5-turbo>

ID	Model	ROUGE F1 _{ref:Cluster}			ROUGE F1 _{ref:GPT-4}			Len	ref:GPT4 BS
		R-1	R-2	R-L	R-1	R-2	R-L		
1	LED	39.41 ± 0.121	23.10 ± 0.149	36.18 ± 0.128	25.93 ± 0.049	3.81 ± 0.029	21.72 ± 0.076	166.0	81.67
2	Flan-T5-XL	<u>38.29</u> ± 0.071	<u>24.31</u> ± 0.048	<u>37.16</u> ± 0.046	30.17 ± 0.149	7.01 ± 0.007	27.24 ± 0.123	148.2	84.86
<i>LLM baselines</i>									
3	LLAMA 2-7B	34.08 ± 0.104	11.73 ± 0.044	32.29 ± 0.084	41.50 ± 0.076	12.28 ± 0.047	36.50 ± 0.059	125.9	88.93
4	LLAMA 2-13B	33.69 ± 0.107	11.47 ± 0.085	31.92 ± 0.106	43.29 ± 0.082	13.33 ± 0.030	38.15 ± 0.134	106.2	89.43
5	Mistral	33.51 ± 0.101	11.72 ± 0.076	31.78 ± 0.043	42.67 ± 0.065	12.52 ± 0.061	37.61 ± 0.060	100.3	89.29
6	GPT-3.5	31.37 ± 0.129	9.41 ± 0.004	29.69 ± 0.064	49.62 ± 0.224	<u>17.19</u> ± 0.190	<u>43.54</u> ± 0.200	98.5	<u>90.40</u>
<i>Self-refine</i>									
7	LLAMA 2-7B	32.06 ± 0.092	9.97 ± 0.113	30.26 ± 0.090	40.99 ± 0.101	11.62 ± 0.020	35.98 ± 0.047	124.2	88.91
8	LLAMA 2-13B	31.36 ± 0.124	9.58 ± 0.071	29.67 ± 0.131	43.06 ± 0.031	12.76 ± 0.069	37.73 ± 0.020	104.2	89.62
9	Mistral	33.40 ± 0.193	11.60 ± 0.146	31.64 ± 0.210	42.53 ± 0.082	12.50 ± 0.013	37.38 ± 0.029	105.6	89.24
10	GPT-3.5	31.24 ± 0.006	9.41 ± 0.016	29.64 ± 0.057	<u>49.62</u> ± 0.016	17.33 ± 0.002	43.58 ± 0.092	99.0	90.41

Table 3: Results of different models on the REFLECTASP. BS refers to BERTScore. For ROUGE F1 scores, results are reported as an average of 3 runs, with $\pm$ standard deviation. The best results for each column are highlighted with **bold**, and the second best results are underlined.

30.26 for the Llama 7B model). Conversely, GPT-3.5 and Mistral did not exhibit significant changes in automatic metrics. To further investigate these outcomes, we conduct in-depth analysis and manual inspections (Sec 6 and Sec 6.2), finding that n-gram-based metrics face challenges in accurately evaluating the quality of summaries. The failures to follow the instructions (Liu et al., 2023b) became one impacting factor in fairly evaluating the quality of summaries. These findings underscore the necessity of delving into the texts to examine the effects of the self-refine framework carefully.

5.2 Human Evaluation

In this section, we conduct human evaluations on the generated summaries. To investigate the effects of self-refinement, we conducted a comparative analysis of summaries generated by both baseline and self-refined versions of LLAMA 2-13B. We additionally include Mistral baseline and GPT-3.5 to evaluate the summary quality of different LLMs. At last, we include GPT-4 to verify the quality of the reference. All systems are anonymized and randomly shuffled to allow for fair evaluation.

We randomly sampled aspect-input pairs and asked two in-house human annotators to evaluate the quality of generated aspect-based summaries. For each summary, annotators are given the original reflections from a lecture and the specific aspect, followed by instructions to evaluate *Relevance to the aspect*, i.e., “Determine if the summaries address the specified aspect exclusively, partially, or not at all.” (Amplayo et al., 2021). One annotator completed annotations for 25 instances, while an-

Model	Rel. to Aspect			Rel. to Aspect <i>Score</i> ↓ (<i>Rank</i>)	R-L Rank	
	Exclu.	Part.	Not		ref:Clu.	ref:GPT-4
<i>Baseline</i>						
LLAMA 2	26.7	71.1	2.2	1.75 (4)	1	2
Mistral	31.1	66.7	2.2	1.71 (2)	2	4
GPT3.5	26.7	73.3	0.0	1.73 (3)	3	1
<i>Self-refine</i>						
LLAMA 2	37.8	62.2	0.0	1.62† (1)	4	3
GPT-4	64.4	35.6	0.0	1.36 (-)	-	-

Table 4: Human evaluation results based on *Relevance to Aspect*, which assesses whether the summary discusses the aspect exclusively, partially, or not at all related. Translated relevance scores and their relative rankings are included for better interpretation. We additionally report the four systems’ relative ROUGE-L F1 rankings from Table 3. A dagger ($\dagger$) indicates that the self-refined system significantly outperforms its baseline, as determined by the Wilcoxon signed-rank test with a p-value < 0.05 . GPT-4 is significantly better than all systems with $p < 0.001$. “-” denotes that the system is not included in the ranking.

other completed 20, with 16 instances overlapped. We performed a reliability test on these double-annotated instances, achieving a Cohen’s Kappa of 0.52 (Cohen, 1960) with moderate agreement.⁴

Results Table 4 reports on the 45 annotated instances. We examine the human evaluation of aspect relevance for all systems and GPT-4 references. Summaries produced by GPT-4 exclusively discuss the given aspect 64.4% of the time, indicating

⁴We include the annotation guideline and additional annotation details in Appendix D.

the promising quality as references. Open-sourced LLMs, including LLAMA 2-13B and Mistral, can generate summaries that are off the aspect (2.2%), and it remains challenging for all baselines to keep the full summary exclusively about the given aspect (over 69% summaries are marked as Partial or None). Additionally, our self-refine approach *significantly* improves the relevance of the aspect (the proportion of summaries that exclusively discuss the aspect increased from 26.7 to 37.8).

We further observe an inconsistency exists between automatic metrics and human evaluation of relevance to aspects. We translate these human-annotated categorical values into numerical scores on a scale of 1 to 3, representing 'exclusive,' 'partially,' and 'not related,' respectively. The average scores and their relative rankings are displayed in the table. The self-refined LLAMA 2-13B model obtains the highest rank while being scored low according to automatic metrics. While baseline models obtain high R-L scores against references, their tendency to include texts not focused on the giving aspect yields lower human ratings on overall aspect relevance. This human evaluation highlights the need for a more nuanced understanding and evaluation of aspect-based summary quality beyond automatic scoring methods.

6 Analysis of Summary Revisions

Human evaluations suggested that summaries become more relevant to the aspect, but how does the self-refine framework help with it? In this section, we present a data-driven study on document-level revisions alongside with LLM suggestions, aiming to understand (1) *Whether suggestions are being incorporated?* and (2) *LLMs use what common strategies in self-refinement.*

6.1 Analysis on Suggestions

We begin by examining the similarity between suggestions and both the original and refined versions, operating under the premise that Large Language Models (LLMs) utilize suggested examples and respond to feedback accordingly. Table 5 presents the absolute differences in ROUGE scores when comparing the baseline summary to the suggestion text and the refined summary to the suggestion text. We find that the refined summary overlaps more with the suggestions for open-sourced LLMs, as ROUGE scores increased by a large margin. We note that the GPT-3.5 model instead has dropped

Model	δ R-1	δ R-2	δ R-L
LLAMA 2-7B	+8.5	+8.5	+4.8
LLAMA 2-13B	+5.2	+5.9	+4.8
Mistral	+2.1	+3.0	+1.9
GPT-3.5	-7.3	-10.7	-7.4

Table 5: Absolute differences in ROUGE scores when compared the summary to suggestion text (Refined summary ROUGE - original summary ROUGE).

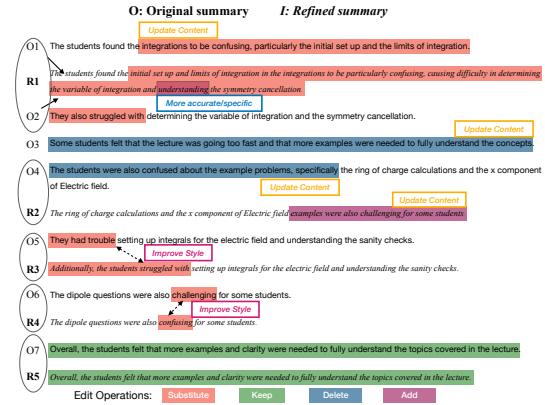


Figure 3: An illustration of the modifications applied through our self-refine method to the original aspect-based summary in Figure 1, using a pipelined system. The pipeline system introduced in Sec 6.2 does the edit extraction and intention label prediction. This revision produces a better aspect-based summary.

similarity towards the suggestions. A closer look at the suggestion texts reveals that the GPT-3.5 models tend to offer more succinct suggestions, and all models present the suggestions in ordered list style (example in Appendix E.1). Since the suggestions are presented in a structured list format (e.g., "1. Emphasize the interactive aspects: *more illustration of the suggestion*"), we examined the text preceding the colon and employed the NLTK toolkit (Bird et al., 2009) to identify the most common n-grams, frequently observed strategies include "concise and focused," "focused on <ASPECT>," "use of more specific language," and "incorporation of specific examples". The subsequent section on edit analysis confirms the validity of these findings.

6.2 Edit Extraction and Intention Identification

To examine the modifications made by the self-refine the approach; we run an automated system (Jiang et al., 2022) to extract edits and determine their underlying intentions. This model is trained on scientific paper revisions, which compromises sentence alignment, edit extraction, and intention

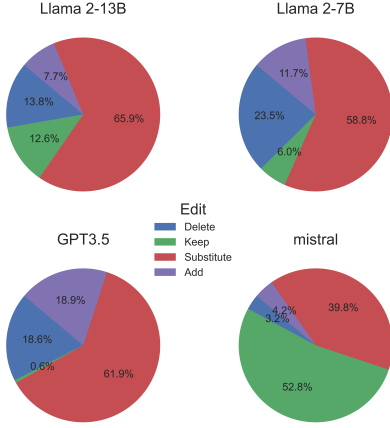


Figure 4: Distribution of edit actions among sentences during self-refinement.

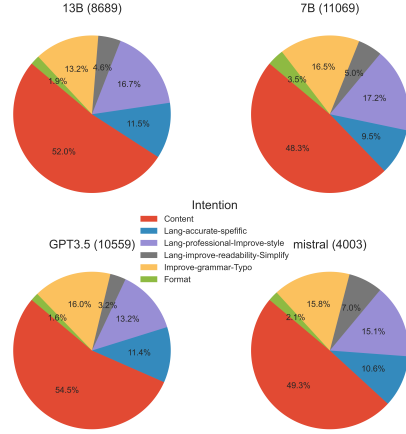


Figure 5: Distribution of span-level edit intentions during self-refinement.

classification modules.⁵ Figure 3 exemplifies modifications made through self-refine.

Analysis of the Updated Sentences We start by exploring the dynamics of sentence-level edit operations, aiming to understand how LLMs modify sentences during self-refinement. As shown in Figure 4, LLMs exhibit different behaviors in their refinement strategies. LLAMA 2 and GPT-3.5 engage in a significantly higher level of substitutions. GPT-3.5, in particular, barely keeps the original sentences, but its modifications do enhance the content’s relevance to the specified aspect, as indicated by the minimal change of automatic evaluations. The LLAMA 2-7B model’s tendency for deletion might explain its decreased performance in automatic metrics by potentially removing valuable content. In contrast, LLAMA 2-13B model keeps a meaningful distribution of different edit operations and was found to benefit from improving the focus on given aspects, as depicted in prior human evaluation. At last, Mistral retains half of the original sentences and favors local changes instead of deleting/adding new sentences.

Analysis of the Edit Intention To understand why the LLM revised the sentences, we run the aforementioned pipelined model on all revised sentences between original and self-refined summaries. The distribution of the intentions is visualized in Figure 5. Most of the edits are classified as content updates. Human evaluation (Sec 5.2) also finds that these updated contents do not necessarily come from the original students’ reflections. Instead, they

often represent suggestions to enhance course instruction, such as "providing more examples and practice questions." This indicates a strategic approach by the LLMs to suggest improvements beyond merely summarizing reflections to offering actionable advice for educational enhancement. Furthermore, language-related edits, especially to *Improve Style* (to make the text sound more professional or coherent without altering the meaning) and *More Specific* (minor adjustment to improve the accuracy or specificity of the description) are prevalent in all models. These findings align with LLM-generated suggestions to “provide specific examples” and “emphasize key points”, highlighting their focus on clarity and precision.⁶

7 Conclusion

In this work, we contribute REFLECTASP, an aspect-based summarization dataset built on top of REFLECTSUMM (Anonymous, 2023), with GPT-4 generated reference summaries. We explore the capabilities of LLMs in the realm of zero-shot open-aspect-based summaries of education opinion text. Our findings reveal that LLMs are capable of generating more focused summaries considering certain aspects of the lecture. With in-depth analysis, our proposed self-refine framework exhibits the possibility of improving the summaries through revision strategies such as elaboration and rephrasing students’ points. Automatic and human evaluations suggested that there still exist gaps between human conception and the machine’s understanding of aspect-based summarization tasks, calling for future work to improve.

⁵Details of the taxonomy and pipeline model are in Appendix E.3.

⁶Described in Section 6.1.

Limitation

This study leverages existing student-written reflections and utilizes the generative power of large language models to produce aspect-based summaries. Although this approach was effective for the specific education dataset we used, it may not be readily applicable to different datasets.

Additionally, while our self-refine approach provides succinct suggestions to improve the summaries from multiple viewpoints, human evaluation indicates that LLM itself may struggle to comprehend and apply the suggestions effectively in revision. This challenge could stem from the inability of smaller models to facilitate the flow of instructions (Skopek et al., 2023; Liu et al., 2023b). We admit that human evaluation is of a small scale, and it can be extended to more diverse dimensions to evaluate the quality of aspect-based summaries.

Moreover, our reliance on conducting experiments in a zero-shot manner may hinder the model from comprehensively understanding the meaning of the prompts and thus fail to produce higher-quality revisions.

Ethical Consideration

Abstractive summarization models have been found to contain hallucinated artifacts that do not faithfully present the source texts. We do not see concerns about applying our model, as user-sensitive information will not be included in the students' reflections.

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, Dublin, Ireland. Association for Computational Linguistics.
- Laila Alrajhi, Ahmed Alamri, Filipe Dwan Pereira, and Alexandra Ioana Cristea. 2021. [Urgency analysis of learners' comments: An automated intervention priority model for mooc](#). In *International Conference on Intelligent Tutoring Systems*.
- 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, Singapore. Association for Computational Linguistics.
- 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Anonymous. 2023. REFLECTSUMM: A benchmark for course reflection summarization.
- Sinem Aslan, Nese Alyuz, Cagri Tanriover, Sinem E. Mete, Eda Okur, Sidney K. D'Mello, and Asli Arslan Esme. 2019. [Investigating the impact of a real-time, multimodal student engagement analytics technology in authentic classrooms](#). In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.
- Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022. [NEWS: A corpus for news topic-focused summarization](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, 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](#).
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- 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, Florence, Italy. Association for Computational Linguistics.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. [RARR: Researching and revising what language models say, using language models](#). In *Proceedings of the 61st Annual Meeting of the*

669	<i>Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 16477–16508, Toronto, Canada.	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback . <i>Neurips</i> .	723
670	Association for Computational Linguistics.		724
671			725
672	Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization . <i>Transactions of the Association for Computational Linguistics</i> , 9:211–225.		726
673			727
674			728
675			729
676			
677	Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text de-generation . In <i>International Conference on Learning Representations</i> .	Muhsin Menekse, Glenda Stump, Stephen Krause, and Michelene Chi. 2011. The effectiveness of students’ daily reflections on learning in an engineering context . pages 22.1451.1–22.1451.10.	730
678			731
679			732
680			733
681	Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet . <i>ArXiv</i> , abs/2310.01798.	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 <i>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)</i> , pages 1080–1089, Online. Association for Computational Linguistics.	734
682			735
683			736
684			737
685			738
686	Jennifer Jacobs, Karla Scornavacco, Charis Harty, Abhijit Suresh, Vivian Jia Yin Lai, and Tamara R. Sumner. 2022. Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change . <i>Teaching and Teacher Education</i> .		739
687			740
688			741
689			742
690			743
691			
692	Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. <i>arXiv preprint arXiv:2310.06825</i> .	Duy-Hung Nguyen, Nguyen Viet Dung Nghiem, Bao-Sinh Nguyen, Dung Tien Tien Le, Shahab Sabahi, Minh-Tien Nguyen, and Hung Le. 2022. Make the most of prior data: A solution for interactive text summarization with preference feedback . In <i>Findings of the Association for Computational Linguistics: NAACL 2022</i> , pages 1919–1930, Seattle, United States. Association for Computational Linguistics.	744
693			745
694			746
695			747
696			748
697	Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process in scientific writing . In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.		749
698			750
699			751
700			
701			752
702			753
703			754
704	Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries . In <i>Text Summarization Branches Out</i> , pages 74–81, Barcelona, Spain. Association for Computational Linguistics.		755
705			756
706			
707			757
708	Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023a. On improving summarization factual consistency from natural language feedback . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.		758
709			759
710			760
711			761
712			762
713			
714			757
715			758
716	Yixin Liu, Alexander R Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2023b. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. <i>arXiv preprint arXiv:2311.09184</i> .		759
717			760
718			761
719			762
720			
721			763
722			764
			765
			766
			767
			768
			769
			770
			771
			772
			773
			774
			775
			776
			777
			778
			779

780	Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 6301–6309, Online. Association for Computational Linguistics.	839
781		840
782		841
783		842
784		
785		
786		
787	Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization . In <i>Proceedings of ACL-08: HLT</i> , pages 308–316, Columbus, Ohio. Association for Computational Linguistics.	
788		
789		
790		
791		
792	Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models .	
793		
794		
795		
796		
797		
798		
799		
800		
801		
802		
803		
804		
805		
806		
807		
808		
809		
810		
811		
812		
813		
814		
815	Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 10837–10851, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	
816		
817		
818		
819		
820		
821		
822	Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct . In <i>The Eleventh International Conference on Learning Representations</i> .	
823		
824		
825		
826		
827	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , pages 38–45, Online. Association for Computational Linguistics.	
828		
829		
830		
831		
832		
833		
834		
835		
836		
837		
838		
	Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. <i>arXiv preprint arXiv:2109.10862</i> .	839
		840
		841
		842
	Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-scale open domain aspect-based summarization . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.	843
		844
		845
		846
		847
		848
		849
	Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative text summarization via ChatGPT . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 10644–10657, Singapore. Association for Computational Linguistics.	850
		851
		852
		853
		854
	Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert . In <i>International Conference on Learning Representations</i> .	855
		856
		857
		858
	Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-hint prompting improves reasoning in large language models. <i>arXiv preprint arXiv:2304.09797</i> .	859
		860
		861
		862
	A Clustering Results	863
	Given the 1,096 aspects, we encoded them using Phrase-BERT (Wang et al., 2021), then applied an unsupervised clustering algorithm, K-means from the scikit-learn package (Pedregosa et al., 2011) ⁷ , to group the phrase embeddings into clusters. The parameters for K-means are { "init": "k-means++", "n_init": 3, "max_iter": 300}. We search for the best N based on the SSE of cosine similarities. Table 6 is one example of clustering results, with 5 aspects per cluster.	864
		865
		866
		867
		868
		869
		870
		871
		872
		873
	B Prompt Templates	874
	B.1 GPT-4 Prompt	875
	We use the prompt in Table 8 for GPT-4.	876
	B.2 Self-refine Prompt	877
	The prompt for our proposed self-refine framework can be found in Table 7.	878
		879
	B.3 Post-processing	880
	We apply a list of rules to clean up the LLM-generated summaries. We remove the first paragraph, which serves as the LLM’s response to the request. We continue by locating the sentence that	881
		882
		883
		884

⁷<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

ID	Cluster Size	Example Aspects
0	32	['excel', 'No Confusion', 'No Confusion', 'No Confusion', 'No Confusion']
1	61	['In-Class Problems', 'In-Class Problems', 'Exam Prep', 'In-class assignments', 'Syllabus', 'Structure of Class']
2	69	['Other Statements', 'Other Statements', 'Other Statements', 'Other Statements', 'Other Statements']
3	85	['Teamwork/Breakout Rooms', 'Capital Investment', 'Groupwork', 'New Project', 'Groupwork']
4	88	['Electric/Uniform Field', 'Energy Calculations and Units', 'Car Carbon Emissions', 'Electric Charges', 'Current/Resistance']
5	108	['Evaluating and citing reliable resources', 'Phone book activity', 'Downloading the file', 'Introduction to the new project', 'Last example question']
6	32	['Assignment 17', 'Assignment 8, A08', 'Assignment 8, A08', 'Assignment 8 OR 5', 'Assignment 13']
7	127	['Redefining Systems', 'Prototyping/Creating Prototypes', 'Engineering Majors', 'Coding', 'Pseudocode and Algorithm']
8	62	['Related to Trig', 'Related to Functions', 'Related to the Quiz', 'Related to the Project', 'Related to Induction']
9	50	['RB BST/Red-Black tree/red-black BST/Red Black BST', 'excel/Excel', '1 vs 100 Sheets Question', 'A10', 'Red Black BSTs']
10	118	['Matlab/matlab/MATLAB', 'Backtracking', 'Porblem scope of the project', 'Breakout Rooms', 'Deck of cards/poker problem']
11	107	['When to use certain graphs', 'Comparing Excel & MatLab', 'Free Body Diagrams', 'Difference between data types (categorical/numerical, nominal/ordinal, discrete/continuous)', 'Histograms']
12	47	['In-Class Demonstrations', 'Meeting People/Professor', 'Videos shown in Class', 'In-Class Demonstrations', 'In-Class Activity or In-Class Assignment']
13	19	['Taum Salk reservoir power activity', 'The Tom Sauk Reservoir', 'Hydropower and Hydroelectricity', 'Hydroelectric dams, power, and reservoirs', 'Taum Sauk Project or Reservoirs']
14	91	['Related to Flux', 'Related to Concepts (Gaussian Surfaces, Faraday Cages, E Fields)', 'Related to Loops', 'Related to Circuits & Graphs', 'Related to Linear Regression']

Table 6: K-Means clustering results of aspects, K =15.

is a conclusive statement of the lecture, such as “Overall, the students expressed ...”. For the self-refine framework, the model may continue generating “suggestions” for improvement after the summary. We opt to exclude those extra contents through regular expression on their patterns. We admit that such processing is not perfect, and the model results are affected by corner cases. We will release all scripts with the datasets.

C Model Implementation Details

All of our experiments are conducted on Quadro RTX 5000 GPUs, each with 16 GB RAM. To tackle the memory limitation and speed up the inference with LLMs, we applied FP4-bit quantization for inference. The 7B models can be run on one card, and the 13B models need two cards for inference.

C.1 Pre-trained Models

LED models are implemented with the Huggingface library (Wolf et al., 2020) using PyTorch, initialized with the “allenai/led-base-16384” check-

point⁸. For different runs, we employ the nucleus sampling method (Holtzman et al., 2020) with p set at 0.9. For Flan-T5-XL models, we experiment with different beam sizes for beam search (3,4,5) to collect different generations across multiple runs.

C.2 LLMs

We employ LLAMA 2-7B/13B-chat (Touvron et al., 2023)⁹, and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)¹⁰ models for experiments. We employ the nucleus sampling method (Holtzman et al., 2020) with p set at 0.9 and a temperature of 0.5. We set the maximum number of tokens to 256 for summary generation in the main tables. We manually evaluated the aspect-based summaries generated during a brief manual tuning of the prompt text to determine the appropriate prompt. The final prompt used for all

⁸<https://huggingface.co/allenai/led-base-16384/tree/main>

⁹<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf> and <https://huggingface.co/meta-llama/Llama-2-13b-chat-hf>

¹⁰<https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1>

Stage	Role	Content
	system:	You are a responsive abstractive summarizer that summarizes the collection of student lecture reflections by focusing on a specific topic.
GENERATION	user:	Please write a short summary with no more than 100 words, focusing on the topic of {topic} based on below reflections: {reflections} SUMMARY:
	sys:	[GENERATED TEXT]
	user	[INST] Can you provide a short list of 2-3 suggestions to improve the generated summary, making it more concise and focused on the topic – topic? The suggestions should be based on the original reflections and generated summaries, don't give generic suggestions. [/INST]
	sys:	[SUGGESTIONS]
Restart the conversation		
	system:	You are a responsive abstractive summarizer that summarizes the collection of student lecture reflections by focusing on a specific topic
REFINE	user:	Please improve the short summary written below, with the suggestions. The revised version should be no more than 100 words, focusing on the topic of {topic} based on below reflections: {reflections}.
		ORIGINAL SUMMARY: {GENERATED TEXT}
		SUGGESTIONS FOR IMPROVEMENT: {SUGGESTIONS}
		REFINED SUMMARY:

Table 7: The self-refine prompt used for open-aspect-based summarization (OABS), with three stages: GENERATION, SUGGESTION, and REFINE.

Role	Content
system:	You are a responsive abstractive summarizer that summarizes the collection of student lecture reflections by focusing on a specific topic.
user:	Please write a short summary with no more than 100 words, focusing on the topic of {topic} based on below reflections:

Table 8: GPT-4 prompt used to generate reference aspect-based summaries.

models is presented in Table 7. We adopt the specific prompt formats according to Meta’s blog (<https://huggingface.co/blog/llama2#how-to-prompt-llama-2>) and the official Mistral repo. For the GPT-3.5 model, we used GPT3.5 turbo 1106 from <https://platform.openai.com/docs/models/gpt-3-5-turbo> as one strong baseline. The temperature is set as 0.5, and the max_token length is set to 256.

C.3 Model Outputs

We include one example of all baseline models, covering LED, FLAN-T5-XL and LLMs in Table 9. LED and FLAN-T5-XL struggle to produce fluent summaries. LLAMA 2-7B generates summaries formatted as ordered lists, whereas LLAMA 2-13B, Mistral and GPT-3.5 models generate outputs in a standard format. The reference from GPT-4 is the most concise and avoids including information unrelated to the topic.

D Human Annotation Details

We conducted evaluations with two PhD students, both of whom have over five years of NLP research experience. The annotators were asked first to read the instruction; then they were given forty randomly sample aspect and input reflection pairs, each accompanied by the generated outputs from the following five models: (1) LLAMA 2-13B – baseline, (2) Mistral – baseline, (3) LLAMA 2-13B – self-refined, (4) GPT-3.5 baseline, and (5) GPT-4. The experts were presented with the aspect, original reflections, and each system output in a unique row of an Excel file. They were then asked to assign a label to evaluate whether the summary discussed the aspect *Exclusively* / *Partially* / *Not at all*. Human annotation guidelines with examples can be found in Fig 7. We also experimented with other criteria such as “Aspect-based Factuality

–Sentence Level”, which aims to annotate each sentence’s factuality based on the human-annotated clusters, yet the annotators’ agreement is too low to be included in the paper. Annotator 1 finished 25, and Annotator 2 annotated 20 instances, with 16 overlapped.

D.1 Confusion Matrix on Relevance Rating

We report the confusion matrix on relevance annotations over the 16 overlapped instances for each system in Figure 6.

E Supplementary Materials about Self-refine Analysis

E.1 Suggestion Examples

In Table 10, we present one sample of suggestions for three systems. Mistral and GPT-3.5 provided shorter suggestions with fewer exemplifications.

E.2 Example of the Self-refine pipeline

We include one example of both pre- and post-self-refine summaries and the corresponding suggestions in Table 11.

E.3 Edit Intention Taxonomy and the Pipeline Model

We adopt the edit intention taxonomy from (Jiang et al., 2022). There are seven fine-grained intention labels:

1. *Improve Language – More Accurate/Specific*: Minor adjustment to improve the accuracy or specificity of the description.
2. *Improve Language – Improve Style*: Make the text sound more professional or coherent without altering the meaning.
3. *Improve Language – Simplify*: Simplify complex concepts or delete redundant content to improve readability.
4. *Improve Language – Other*: Other language improvements that don’t fall into the above categories.
5. *Correct Grammar/Typo*: Fix grammatical errors, correct typos, or smooth out grammar needed by other changes.
6. *Update Content*: Update large amount of scientific content, add or delete major fact.

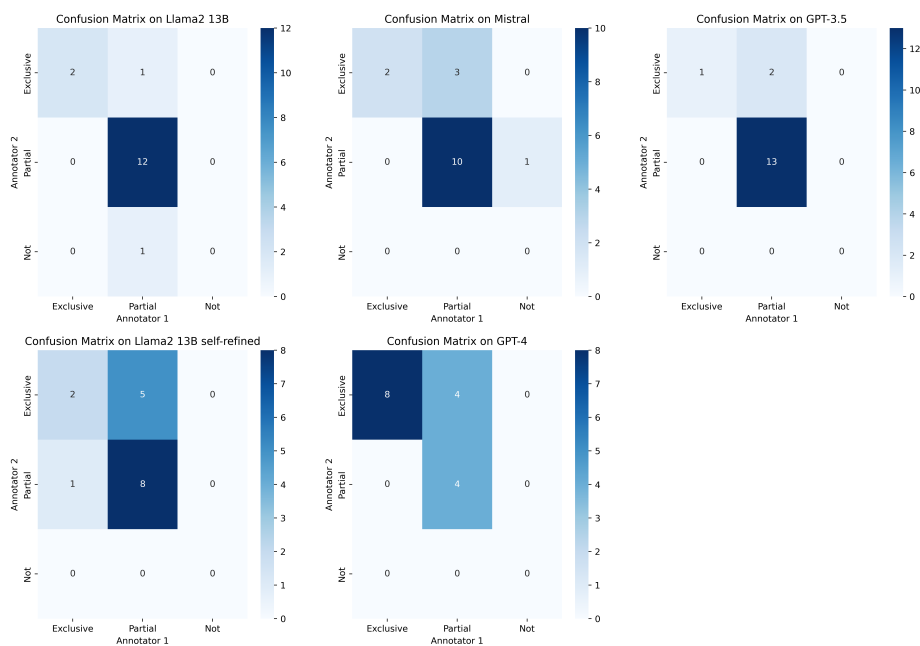


Figure 6: Confusion matrix of human evaluations on *Relevance of Aspect*.

7. Adjust Format: Adjust table, figure, equation, reference, citation and punctuation etc.

We recommend the reader checking the original paper for more details.

Pipelined Model We run the pipeline script from <https://github.com/chaojiang06/arXivEdits/tree/main/code/pipeline> to predict edit and intention results.

Your task involves assessing the quality of five summaries generated from a set of student reflections. These reflections were prompted by professors who requested students to write about points of confusion or interest. The summaries are generated to condense students' opinions, focusing on specific aspects. These aspects can include course topics such as "Gaussian Surface" or "Assignments."

To perform this evaluation effectively, please adhere to the following steps:

1. **Check the aspect that is used to prompt the model for summary generation** (aspects here can be course terms such as "Gaussian Surface" or "Assignments")
2. **Review the system summaries**
3. For each summary, you will need to evaluate the **Relevance to the aspect**: *Determine if the summaries address the specified aspect exclusively, partially, or not at all.*

Example

Input Reflections

(highlighted reflections are manually annotated as related to the given aspect)

- i think the last practice problem we did in class could've been done slower to make it easier to comprehend. i had to think it over for a few minutes after class ended for it to come to me
- I did not understand why the electric field inside of a solid, uniform, insulating sphere ($r < R$) is greater than the field outside of it ($r > R$).
- I think the angle at which flux flows from an object is a bit confusing. The first problem on the concept quiz deals with this.
- I think the idea and concepts make sense, but I am confused about what shapes to choose as the Gaussian surface. Can you choose any shape? Are there multiple shapes that will still give the same answer as the cylinder? Also, the conductor vs insulator in reference to Gauss' law is confusing to me
- The orientation of planar electric fields and cylindrical Gaussian surface
- More examples would be helpful for further understanding.
- Figuring out what Gaussian surface to draw in problems
- I was not too confused but the field lines of the right angle planes question for the quiz was confusing.
- Nothing was too confusing. The only part I found a little difficult was the different gauss fields around the uniformly charged inner sphere. Just transitioning from one layer to the next for me slightly confused in your last example problem video.
- I was confused about why we find the areas of the three different parts of the cylinder, and how we do this.
- The second question on the concept quiz confused me, I wasn't sure how positive and negative charges being outside or inside affect flux.
- Are these the only three symmetries we will work with?

All system summaries are generated with the prompt

```
``Please write a short summary with no more than 100 words, focusing on the topic of {ASPECT} based on below reflections``
```

Write a caption...

Example 1: (Relevance to the aspect)

Whether the summaries discussed the given aspect *exclusively, partially, or not at all.*

Aspect: Gaussian Surface

Generated Summary:

The students found the concept of Gaussian surface to be challenging, particularly the choice of the surface and its relationship to the electric field. They also struggled with the idea of flux and how it is affected by positive and negative charges being inside or outside the surface. Additionally, the students found the orientation of planar electric fields and cylindrical Gaussian surface to be confusing.

To address these challenges, the students suggested more examples and practice problems to help them better understand the concept of Gaussian surface and its application to electric fields. They also recommended that the instructor provide more clarification on the choice of the surface and how it affects the electric field.

Overall, the students found the topic of Gaussian surface to be confusing, but with more practice and clarification, they were able to understand the basic concepts.

Relevance to the aspect: Partially

(Rationale): The first paragraph discusses students' reflections on Gaussian Surfaces. However, sentences highlighted in yellow are unrelated to Gaussian Surfaces, thus the overall relevance is **Partially**.

Figure 7: Human annotation guideline.

Model	Summary
LED	The conservation of energy question was very tricky. I was indecisive between the correct answer and the difference in charge answer, so I felt that I learned the most from this question. The explanation of your plastic sheet and lightbulb experiment was really neat and caught my interests. I thought having the visual examples was interesting because it makes topics easier to understand. I found it interesting how similar electric force is to gravitational force. I found the demonstration done to show the lightbulb lighting was very interesting because of the separation of charges. The group work in solving the distribution of charge and electric force. What I found most interesting today was that there can be two components of an electrostatic force. I found this interesting because I'm working with forces that are invisible. Energy conservation in particle physics with a gamma ray photon being split into an electron and positron was very intriguing since I've never learned much about gamma rays
FLAN-T5-XL	I thought the demonstration with the lightbulb was fun to watch. It gave me a visual representation of what was happening. I found interesting how similar the forces caused by charges and gravity are. I like that the math is pretty consistent and the laws and theorems are pretty easily definable I just get messed up when you're doing demonstrations that move really quickly and the words go over my head
LLAMA 2-7B	Many students found the visual demonstrations of the lightbulb experiment to be interesting and helpful in understanding the concepts of electric forces and charge distribution. * The comparison of electric forces to gravitational forces was seen as interesting and helpful in understanding the nature of these forces. * The use of visual aids, such as the plastic sheet and lightbulb experiment, was found to be effective in making complex topics more accessible. * The group work in solving the distribution of charge was seen as interesting and helpful in understanding the collaborative nature of scientific problem-solving. * The demonstration of the lightbulb lighting was found to be interesting and helpful in understanding the concept of separation of charges.
LLAMA 2-13B	The students found the demonstrations and interactive aspects of the lecture to be the most interesting, including the plastic sheet and lightbulb experiment, the split-screen video, and the group work in solving the distribution of charge. They also appreciated the visual representations of concepts, such as the lightbulb demonstration, as it made the topics easier to understand. The similarities between electric force and gravitational force were also a source of interest. Additionally, the students found the concept of electric charge properties and the different ways charges spread out to be fascinating. Overall, the students enjoyed the interactive and visual aspects of the lecture and found the topic of Light Bulb/Demonstrations to be the most engaging.
MISTRAL	The students found the demonstration with the lightbulb to be visually interesting and helpful in understanding the topic. They also found the concept of electric charge properties and the distribution of charge to be interesting. The students appreciated the opportunity to interact with the instructor and engage in group work. The students found the explanation of the plastic sheet and lightbulb experiment to be neat and caught their interests. They also found the conservation of energy question to be challenging and learned the most from it. The students found the concept of energy conservation in particle physics with gamma rays intriguing and the clicker question about F and F' to be interesting.
GPT-3.5	The students found the light bulb demonstration and other visual examples interesting and helpful in understanding electric charges. They also appreciated the interaction with the instructor and the breakout room activities. The similarity between electric and gravitational forces intrigued them. They found the conservation of energy question and the demonstration of the light bulb lighting particularly tricky and interesting. Additionally, they found the concept of charges spreading out and the shell theory to be engaging. The practical application of concepts through demonstrations and practice problems was well-received
GPT-4Reference	Students found the lightbulb demonstrations engaging, providing a clear visual representation of abstract concepts like charge separation and electrostatic forces. The hands-on experiment with the plastic sheet and lightbulb piqued interest, making the topic more accessible. The visual aspect of seeing the lightbulb light up during the demonstration was highlighted as particularly interesting, enhancing understanding of the principles being taught.

Table 9: An example of different system summaries. The aspect is “Light Bulb/Demonstration” .

Model	Suggestion
LLAMA 2-13B	Sure, here are three suggestions to improve the generated summary, making it more concise and focused on the topic – Light Bulb/Demonstrations, based on the original reflections and generated summaries: 1. Emphasize the interactive aspects: The students found the interactive aspects of the lecture, such as the demonstrations and group work, to be the most interesting. Therefore, the summary could be improved by highlighting these aspects and how they contributed to the students’ engagement and understanding of the topic. 2. Focus on the visual representations: The students appreciated the visual representations of concepts, such as the lightbulb demonstration, as it made the topics easier to understand. The summary could be improved by emphasizing the importance of visual aids in the lecture and how they helped the students grasp the concepts better. 3. Highlight the similarities between electric force and gravitational force: Many students found the similarities between electric force and gravitational force to be interesting. The summary could be improved by highlighting this aspect and exploring the implications of these similarities in the context of Light Bulb/Demonstrations.
MISTRAL	1. Focus on the most interesting and relevant aspects of the reflections related to light bulb/demonstrations, such as the visual appeal of the light bulb demonstration and the concept of electric charge properties. 2. Use specific examples from the reflections to illustrate the students’ engagement and understanding of the topic, such as the conservation of energy question and the shell theory.3. Use more concise language and avoid repeating information from the reflections, such as the students’ appreciation for the instructor and group work.
GPT-3.5	1. Focus exclusively on the students’ confusion regarding RC circuits and related concepts, omitting references to other topics such as magnetic fields and matrices. 2. Highlight the specific challenges students faced with RC circuits, such as understanding the loop rule and solving multi-loop circuits, while omitting general statements about finding concepts clear. 3. Provide more specific examples of the areas in RC circuits where students struggled, such as determining currents through resistors and the relationship of time constant to voltage and current.

Table 10: Suggestion examples of different systems.

Stage	Output
GENERATION	The students found the integrations to be confusing, particularly the initial set up and the limits of integration. They also struggled with determining the variable of integration and the symmetry cancellation. Some students felt that the lecture was going too fast and that more examples were needed to fully understand the concepts. The students were also confused about the example problems, specifically the ring of charge calculations and the x component of Electric field. They had trouble setting up integrals for the electric field and understanding the sanity checks. The dipole questions were also challenging for some students. Overall, the students felt that more examples and clarity were needed to fully understand the topics covered in the lecture.
SUGGESTION	Sure! Based on the student lecture reflections and the generated summary, here are 2-3 suggestions to improve the summary, making it more concise and focused on the topic of Integration: 1. Emphasize the specific topics that caused confusion: In the generated summary, the students' confusion is mentioned in general terms. To make the summary more focused on Integration, the specific topics that caused confusion, such as the initial set up, limits of integration, and determining the variable of integration, should be emphasized. 2. Highlight the most common sources of confusion: The students' reflections mention various topics that caused confusion, but some are more common than others. To make the summary more concise, the most common sources of confusion, such as the limits of integration and determining the variable of integration, should be highlighted. 3. Use more specific language: The generated summary uses general language to describe the students' confusion. To make the summary more specific and focused on Integration, more technical language should be used to describe the concepts that caused confusion, such as "limits of integration" and "determining the variable of integration".
REFINE	The students found the initial set up and limits of integration in the integrations to be particularly confusing, causing difficulty in determining the variable of integration and understanding the symmetry cancellation. The ring of charge calculations and the x component of Electric field examples were also challenging for some students. Additionally, the students struggled with setting up integrals for the electric field and understanding the sanity checks. The dipole questions were also confusing for some students. Overall, the students felt that more examples and clarity were needed to fully understand the topics covered in the lecture.

Table 11: The self-refine example on LLAMA 2-13B. Aspect is "Integration".