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 novel dataset that summarizes student reflections on STEM lectures. Despite the promising performance of large language models in general summarization, their application to nuanced aspect-based summaries remains under-explored. ReflectASP eases the exploration of open-aspect-based summarization (OABS), overcoming the limitations of current datasets and comes with ample human annotations. We benchmarked different types of zero-shot summarization methods and proposed two refinement methods to improve summaries, supported by both automatic and human manual evaluations. Additionally, we analyzed suggestions and revisions made during the refinement process, offering a fine-grained study of the editing strategies employed by these methods. We will make our models, 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) focuses on shareable subtopics across documents, such as summaries of customer reviews emphasizing the *room* and *location* aspects of hotels (Angelidis et al., 2021a), the emergence of open-aspect-based summarization (OABS) (Tan et al., 2020; Yang et al., 2023; Amar et al., 2023) enables the identification of unique aspects for each document and the generation of tailored summaries.

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; Menekse, 2020; Kim, 2024) 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, 26 students wrote reflections after a physics lecture, which covered different aspects such as *Grounding*, *Current* and *Circuit*, etc. While a prior corpus in this domain (REFLECTSUMM (Zhong et al., 2024)) provides human-written generic summaries, when focusing on a specific aspect such as *Grounding*, this generic summary merely notes that *"these aspects were found interesting"*. It thus fails to provide instructors with meaningful insights into how students engage with these aspects. For example, details like *"they enjoyed how grounding simplifies calculations"* would help instructors better prepare for future lectures and ensure that students are following the material effectively. Similarly, for reflections on confusing points, the aspect-based summaries can pinpoint the major sources of confusion. This allows instructors to revisit these topics or provide additional practice examples. Additional examples contrasting generic versus aspect-based reflection summaries are provided in Appendix A.

Student reflections offer a robust domain for OABS research due to their inherent complexity. The **aspects** can vary across disciplines, e.g., "sorting algorithms" is highly relevant to Computer Science while "in class activities" is relevant across course disciplines. This diversity tests the summarizing model's capability in capturing aspect-specific information within different contexts. Furthermore, students may articulate their reflections differently, even on the same aspect. The model must discern and synthesize both the shared underlying challenges and the distinctive insights unique to each student's perspective. Additionally, evaluating and improving the capabilities of large language models (LLMs) for generating aspect-based summaries in the educational domain is strategi-

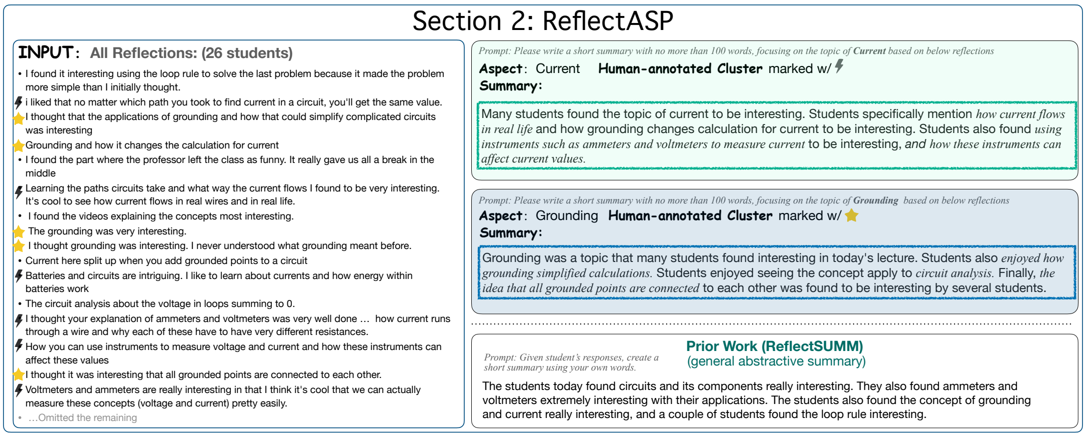

Figure 1: An example from our ReflectASP dataset. For a given collection of reflections and a specified aspect, we provide human-annotated clusters of aspect-related reflections (highlighted in the input) and human-written aspect-based reference summaries (top right). Unlike prior work (REFLECTSUMM), which produces a single generic summary (bottom right), our annotated dataset includes individual summaries for each aspect, accounting for cases where multiple aspects exist within a single input. These aspect-based summaries capture more detailed elements of students' reflections (*italicized*) such as *enjoyed how grounding simplified calculations* in the second summary, offering valuable insights that can help instructors enhance their teaching.

cally critical. Unlike other text domains potentially compromised by training data leakage (Zhou et al., 2023), the educational data, collected from real-world scenarios and excluded from web sources, offer a more rigorous assessment of model performance. Once more capable LLMs are validated on the task, they can be deployed to real-world learning systems, helping instructors improve lectures and providing platforms for peer learning.

We introduce Reflective ASPect-based summarization (ReflectASP), a novel dataset containing 313 manually annotated data instances, including aspect, source student reflections, annotated aspect-based clusters of student reflections, and human-written abstractive aspect-based summaries. Our dataset addresses the lack of open aspect-based summarization resources in education. Built on real-world application data, it avoids potential data contamination and provides a fairer evaluation and is of similar scale when compared to the test split of similar manually curated datsets (Amplayo et al., 2021; Takeshita et al., 2024), thus suffice for robust performance validation. It also features human-annotated abstractive summaries, ensuring natural and coherent text, as well as high-quality annotations of supporting aspect clusters, providing further validation of summary quality.

Using this corpus, we benchmark various zero-shot summarization approaches and propose two refinement methods that leverage LLMs' capability to self-critic and improve (Madaan et al., 2023; Huang et al., 2023; Welleck et al., 2023). Our experiments, covering multiple LLM backbones, include diverse automatic evaluations and human evaluations to validate the benefits of each approach and provide suggestions for future work in this novel domain and related aspect-based summarization tasks. Finally, we conduct a data-driven analysis of both the refinement suggestions and pre-/post-refinement summaries to identify common strategies used by LLMs to improve aspect-based summary generation, shedding insights for future work in the aspect-based summarization task.

2 Constructing the REFLECTASP Corpus

Dataset Curation. The student reflections in REFLECTASP are a subset of those in REFLECTSUMM (Zhong et al., 2024), which comes with phrase-based, extractive and generic abstractive summaries. For each lecture, the dataset provides a collection of student reflections focusing on interesting or confusing points. Annotators are directed to extract five noun-phrases summarizing the reflections, and mark original student reflections as evidence for their annotated noun phrases. Out of the 782 reflections-summary pairs in the dataset, we construct our dataset by treating all reflections as the multi-document summarization input 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

	Domain	Collect.	Sum. Rew.	Incl. Ext	# Test Set	$\textbf{Word}_{\text{input}}$	$\textbf{Doc}_{\text{input}}$	Aspect	$\textbf{Word}_{\text{sum.}}$	Novelty-n(1/2/3)	Comp. Ratio
FACETSUM	Scientific	A	No	✗	6,000	6,827	-	4	290	-	-
ASPECTNEWS	News	M	No	✗	400	248	-	4	115	-	-
SPACE	Reviews	M	Yes	✓	150	14,335	100	6	26	0.02/0.23/0.50	704:1
OPOSUM+	Reviews	M	Yes	✓	120	1,002	10	18	30	0.11/0.49/0.69	30:1
ACLSUM	Scientific	M	Yes	✓	300	915	-	3	22	0.16/0.58/0.76	41:1
OASUM	Wikipedia	A	No	✗	112,005	1,612	-	-	40	-	-
OPENASP	News	M	No	✗	596	6,860	26	576	82	0.11/0.49/0.74	68:1
LEXABSUMM	Legal	A	No	✗	148	14,357	-	50	251	0.07/0.49/0.70	66:1
REFLECTASP (ours)	Education	**M**	**Yes**	✔	313	817	43	280	69	0.19/0.63/0.84	12:1

Table 1: Descriptive statistics comparing prior datasets (top) to REFLECTASP with their test split. The first five are on ABS, and the others belong to OABS. For the **Collection** method of aspect-based summaries, **A** denotes "Automatic" and **M** denotes "Manual." We distinguish human-rewritten aspect-based summaries (**Sum. Rew.**) from those extracted from the input or generic summaries. **Incl. Ext.** refers to human-annotated content extracted from the input to support the abstractive summary. **# Test Set** is the number of instances in the test split (document set + aspect + summary). $\textbf{Doc}_{\text{input}}$ measures the average number of input reflections/documents/articles. We also report the number of unique aspects (**Aspect**), input length ($\textbf{Word}_{\text{input}}$) and summary length ($\textbf{Word}_{\text{sum.}}$) in words. We also report the proportion of novel n-grams absent from the input (**Novelty-n**) and the input-to-summary length ratio (**Comp. Ratio**). The dash (-) indicates that the metric is not applicable. Gray rows are comparable to our corpus given the **bold** features in the last row.

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, which was further reduced to 1064 by removing phrases on "No Confusion". There exists 767 distinct aspects, highlighting the open-aspect nature of the dataset compared to other corpora. Our further 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". We include details in Appendix B.

Gold Reference Summaries. We recruit two in-house annotators to annotate a subset of 313 unique aspect-lecture pairs. Annotators were first trained in two batches to understand and grasp the tasks before beginning assigned real jobs. We explored two approaches in constructing the summaries: (1) clustering all reflections and drafting the summary from scratch and (2) extracting aspect-related input and revising on top of a GPT-4 generated summary. A pilot study on ten samples suggested that the second option retained good quality through manual inspection and significantly reduced the time needed to write the summary (from 40 mins to 15 mins per data point). We thus apply the second option to produce the full corpus, with full details in Appendix C. We measure inter-annotator performance in ROUGE (Lin, 2004) (R-1/R-2/R-L), which are 48.1/21.9/35.1 among 90 doubly-annotated instances (of the 313 instances).

Dataset Analysis. We discuss properties of REFLECTASP that emphasize its underlying diversity from several angles. The input document lengths vary from 39 to 2467 tokens (Figure 6 in the Appendix), averaging 817 words. The summary length ranges from 16 to 145 tokens, with a median input-to-output (compression) ratio of 12:1. The aspect label length ranges from 1 to 8 words, showcasing the diversity of aspects. The measurement on novelty-n (See et al., 2017) further confirmed that summaries contain a certain level of attractiveness by using new words not present in the input (0.19/0.63/0.84 for 1/2/3 grams respectively). Overall, REFLECTASP requires models to perform well on abstractive forms of summarization. Details for these analyses are in Appendix D.

Comparison to Other Datasets. Table 1 compares our REFLECTASP to existing ABS corpora, including FACETSUM (Meng et al., 2021), ASPECTNEWS (Ahuja et al., 2022), SPACE (Angelidis et al., 2021a), OPOSUM+ (Amplayo et al., 2021), and ACLSUM (Takeshita et al., 2024), as well as to OABS corpora, including OASUM (Yang et al., 2023), OPENASP (Amar et al., 2023), and LEXABSUMM (T.y.s.s. et al., 2024). We highlighted the datasets that are comparable to ours, focusing on the key features of *manual collection*, *human-rewritten* summaries, and *including annotated extractive supporting sentences*. Among corpora with human-rewritten summaries (which guarantees the coherence of the reference), our dataset's size is comparable or larger, making it sufficient for evaluation purposes. In the domain of OABS, our dataset is the only one that includes high-quality human-rewritten summaries along with supporting annotations on the source side. In contrast, OpenASP bypasses the source document, relying instead on manually extracting portions of generic summaries. This approach not only limits its quality to that of the original summaries but also risks compromising text coherence when extracting content from different locations. Additionally, unlike prior work that often relies on extreme compression (e.g., SPACE compressing 14k words into summaries averaging 26 words), our dataset strikes a balance between quality and abstractiveness. Our dataset uniquely focuses on the under-explored educational domain, offering potential real-world applications to enhance teaching performance.

3 Aspect-based Summarization Task

Given course reflections from one lecture and an aspect such as "Integration", we experiment with large language models and test how well they can pick up the salient reflections to generate an abstractive aspect-based summary of the reflections.

LLM Backbones. For LLM backbones, we selected different versions of powerful open-sourced models: LLAMA 3-8B (LLAMA3), LLAMA 3.1-8B (LLAMA3.1) and also LLAMA 3.1-70B (Dubey et al., 2024), all with instructed version. We further add proprietary LLMs GPT-3.5, GPT-4 and GPT-4o as strong baselines. Implementation details are in Appendix E.[1]

Methods. Given different backbone LLMs, we instruct the models using combinations of different methods: (1) **Baseline** uses a basic prompt (see Table 8 in Appendix F). (2) **Self-Refine** uses a *Generate-Suggest-Refine* framework to use the model to improve its outputs (details and prompt in Appendix F.2). This design aligns with the prompt-chaining in Sun et al. (2024), which was proven to be effective. (3) **DCR** (Wadhwa et al., 2024) employs a *Detect-Critique-Refine* pipeline with models finetuned for each phase. We used their released models based on LLAMA3. (4) **E2A** (mimicking the human annotation instructions) uses an extract-then-abstract approach (Takeshita et al., 2024) by prompting the model to first extract relevant student reflections, then generate the abstractive summary. (5) **E2A w/ MC-Refine** harnesses a fact checker to help identify the salient errors in the generated summaries and refine accordingly. Given the initial summary generated from E2A, we apply MINICHECK (MC) (Tang et al., 2024) to evaluate factuality of individual sentences, utilizing the system-extracted aspect-relevant reflections. Next we generate sentence-level error detection and revision suggestions among those detected sentences. However, instead of relying on fine-tuned critique and feedback module, we instruct the LLM to detect the spans within the sentences and provide revision suggestions accordingly. In the end, the LLM is prompted to incorporate all sentence-level revision suggestions to generate the final refined version. We visualize different methods in Figure 2 and include prompts in Appendix F.

Evaluation Metrics. Given the gold references, we measure **ROUGE** F1s (Lin, 2004) (ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (R-L)) and **BERTScore** F1 (Zhang* et al., 2020) (BS). To assess factual accuracy, we report the proportion of summary sentences supported by the documents using the SOTA fact-checker MINICHECK (Tang et al., 2024). We report $\mathbf{MC_{EXT}}$ and $\mathbf{MC_{INPUT}}$, which evaluate factuality based on annotated aspect-related reflections and the full lecture reflections as grounding documents, respectively. Metric details can be found in Appendix G.

4 Results

This section addresses two research questions: **RQ1**. How well do LLMs generate aspect-based summaries of reflections in a zero-shot setting? **RQ2**. How does different refinement help with the summarization? We conducted automatic and human evaluations to validate our findings.

4.1 Automatic Reference-based Evaluation

RQ1. Table 2 shows models' performance with different LLM backbones. Comparing among the baseline prompt method (lines 1, 7, 13), we observe that stronger and more advanced models generally obtained higher R1, R-L and BS. This is also evident in the factuality evaluation scores (MC_{EXT} and MC_{INPUT}). We also note that proprietary LLMs obtain lower automatic scores than the open-sourced LLMs, indicating that their wording might not align with human-written references (rows 19-21). Different from the findings in ACLSUM(Takeshita et al., 2024), we observe that *E2A can help con-*

[1]We include additional results for two weaker backbones (LLAMA 2-13B-chat (Touvron et al., 2023) and MISTRAL (Mistral-Nemo) (Jiang et al., 2023)) in the Appendix.

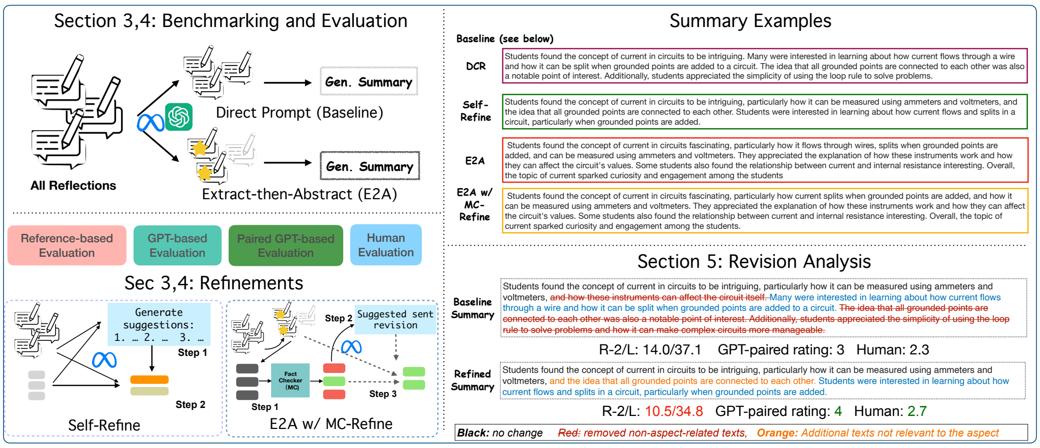

Figure 2: Left: the exemplification of different approaches on leveraging LLMs in zero-shot open-aspect-based summarization task (§3 and §4). Top Right: the outputs of different refining approaches (*DCR*and *Self-Refine*), as well as our proposed approaches (*E2A* and *E2A w/ MC-Refine*). Bottom Right: an analysis of the revision process, showing the evaluation discrepancies among reference-based evaluation, GPT-based evaluation, and human evaluation.

solidate the aspect-relevant information. We posit that this is due to the differing compression ratios of the summaries: ours are longer and include aspect-specific details, making them sufficient to evaluate the behavior of different approaches, unlike ACLSUM extremely short, one-sentence summaries. Among different variants of LLAMA3 and LLAMA3.1 models (row 1 vs. 4, row 7 vs. 10, and row 13 vs. 16), the E2A approach obtained significant improvements across different metrics. Such gain is more salient in smaller models such as LLAMA3. Additional results are in Appendix H.

RQ2. We first assess the revision effects of the Self-Refine and DCR methods (the second and third rows in each block) across all three LLM backbones. Overall, both obtained worse ROUGE and BERTScore results compared to human-written references. However, while the performance gap for Self-Refine narrows as the model improves (rows 1/2 vs. rows 13/14), DCR exhibits greater performance degradation. The analysis of generated summary length (last two columns in Table 2) shows that DCR tends to aggressively shorten the original summary, likely due to the domain mismatch between its training data (meeting summarization) and our education dataset. [2] Our proposed *E2A w/ MC-Refine* overall generate better summaries compared to the baseline prompting on Llama3.1 and LLama3.1-70B (rows 12/18 vs. 7/13). Moreover, it obtained highest factuality scores on LLAMA3.1-70B (row 16 vs. 18) through revision. We attribute it to the strong extractive capability of the LLM (LLAMA3.1-70B obtains 79.6 R-L in extracting the supporting aspect-based reflections, comparable to human extracting performance).[3]

Prior studies have shown that some models, despite achieving higher human ratings, may underperform on reference-based metrics (Zhang et al., 2023; Wadhwa et al., 2024). This motivates us to pursue further evaluations, incorporating both human assessments and analyses assisted by large language models (LLMs) in the following section.

4.2 Automatic GPT-based Evaluation

To overcome the drawbacks of lexical-based metrics, following Wadhwa et al. (2024), we included multiple GPT-4 based evaluation metrics: **GPT-4 Factuality Likert Scale Score (G)**, which uses GPT-4 to score generations when provided a well-defined rubric (Li et al., 2024). We scored the refined/new generation and the initial/baseline output individually and also report the score differences **ΔG**. Additionally, we compute the the pairwise score difference (**Pair ΔG**) between refined/new generations and initial/baseline outputs, and use them to determine the fractions of Wins (*W*), Same Scores (*S*), and Losses (*L*). The order of responses is randomized during the evaluation. Scoring prompts and metric details are in Appendix G.

We report the GPT-based evaluation results for LLAMA3 and LLAMA3.1-70B in Table 3 to compare between the weakest and strongest open-

[2]The differences in content length impact MC_{EXT} and MC_{INPUT} so we omit DCR's MC scores in the table.

[3]We include detailed analysis in Appendix H.3.

ID	Model	R-1	R-2	R-L	BS	MC_{EXT}	MC_{INPUT}	# Sents	# Words
1	LLAMA3	45.99	18.32	41.29	89.89	39.31	83.04	4.15	105.8
2	*w/ Self-Refine*	44.16	15.87	39.03	89.58	36.87	79.45	3.28	88.6
3	*w/ DCR*	38.67	13.66	34.88	89.68	56.61	85.49	2.74	47.2
4	**E2A**	**48.43***	**19.51***	**42.83***	**90.22***	**56.59***	**89.35***	3.12	74.3
5	*w/ DCR*	35.02	14.13	31.53	89.70	71.17	91.27	1.86	30.9
6	*w/ MC-Refine*	44.69	17.02	41.88	90.01	53.87*	87.59*	2.48	60.9
7	LLAMA3.1	46.21	17.57	41.02	89.89	40.72	85.11	3.49	94.8
8	*w/ Self-Refine*	43.68	15.29	38.41	89.53	38.98	79.10	3.18	90.1
9	*w/ DCR*	37.95	13.70	34.09	89.75	56.74	86.38	2.39	42.5
10	**E2A**	**48.08***	**18.97***	**42.78***	**90.12***	**57.79***	**90.44***	3.32	78.7
11	*w/ DCR*	34.97	13.78	31.59	89.70	72.70	92.60	1.94	32.4
12	*w/ MC-Refine*	46.99*	18.13*	41.88*	89.16	57.35*	88.85*	3.25	75.9
13	LLAMA3.1-70B	47.54	18.64	42.12	90.04	53.97	89.48	3.55	94.9
14	*w/ Self-Refine*	46.05	16.88	40.37	89.90	55.74*	86.58	2.93	82.6
15	*w/ DCR*	38.25	14.08	34.22	89.89	71.08	91.77	2.45	42.2
16	**E2A**	**48.75***	**18.79***	**43.17***	90.08	69.55*	88.20	3.51	84.7
17	*w/ DCR*	34.84	13.47	31.25	89.83	82.43	93.37	1.94	32.1
18	*w/ MC-Refine*	47.80	18.44	42.40	**90.34***	**71.79***	**90.50***	3.43	79.9
Proprietary LLMs									
19	GPT3.5-turbo	35.72	9.95	32.05	88.04	44.85	91.25	4.96	100.8
20	GPT4	33.51	7.08	29.06	87.50	51.30	89.82	4.05	97.3
21	GPT4o	34.15	7.93	30.48	87.52	58.07	91.57	4.89	102.9
32	Human (Oracle)	*N/A*				76.97	87.42	3.49	69.2

Table 2: Experimental results on REFLECTASP. All results are averaged over three runs. Gray rows indicate the baseline models, and * means the score is significantly better than the baseline models within each block. The best score for each backbone are **bold**. Light colored cells are not directly comparable to other cells.

sourced models.[4] A more complete table can be found in Appendix H.2. Regarding RQ1, similar to the findings in §4.1, according to the scoring rubric, LLMs can generate summaries that are "overall factually consistent, with a few inconsistencies with the source materials" (rounded to 3), and E2A approaches can further improve. On RQ2, unlike the trend observed in reference-based methods, all refined approaches are found to bring performance gains when compared to the baseline with simple prompt (as evidenced by the positive values of Pair ΔG and Win rate W). These improvements are more profound in the smaller model. While DCR tends to truncate contents more aggressively, it enhances the factuality of generated summaries, as indicated by the significant value on Pair ΔG and Win rate. However, it is also worth noting on the larger Lose rate compared to other refinement approaches, suggesting that this approach can not constantly improve the summary's quality. Our proposed E2A w/ MC-Refine approach obtained the highest GPT-4 rating, and pair-wise test (last block) indicated sigificant improvements.

Model	G↑	ΔG↑	Pair ΔG↑	W↑	S	L
Pairwise Comparison with Baseline Summary as the Original Input						
LLAMA3	2.74	-	-	-	-	-
w/ Self-Refine	2.74	0.00	0.10*	0.19	0.71	0.10
w/ DCR	2.78	0.02	0.25*	0.40	0.46	0.14
E2A	$2.85^{\dagger}$	0.11*	0.28*	0.35	0.57	0.08
E2A *w/ MC-Refine*	**2.87**	**0.11***	**0.31***	0.35	0.59	0.06
LLAMA3.1-70B	2.85	-	-	-	-	-
w/ Self-Refine	2.84	-0.04	0.14*	0.24	0.65	0.11
w/ DCR	2.88	0.03*	**0.17***	0.39	0.42	0.20
E2A	2.87	0.02	0.11*	0.19	0.73	0.09
E2A *w/ MC-Refine*	$\mathbf{2.91}^{\dagger}$	**0.04***	0.14*	0.20	0.73	0.07
Pairwise Comparison with E2A Summary as the Original Input						
LLAMA3 E2A	2.85	-	-	-	-	-
w/ MC-Refine	2.87	0.02	0.10*	0.18	0.74	0.08
LLAMA3.1-70B E2A	2.87	-	-	-	-	-
w/ MC-Refine	$2.91^{\dagger}$	0.02	0.10*	0.13	0.84	0.03

Table 3: GPT-related evaluation results of different methods. Within each block, pairwise metrics compare the outputs of the given system to those in the highlighted rows. A † indicates significant improvement over the previous row (p<0.05) based on a paired bootstrap test, while * denotes that the absolute value is significantly different from zero.

4.3 Human Evaluation

We conduct human evaluations on the generated summaries. Using Amazon Mechanical Turk, we randomly selected 50 samples from ReflectASP and collect annotations on summaries generated by ten different systems. For each sample, three annotations were obtained (thus in total of 1500 annotations), with document-level metrics (Relevance to Aspect and Consistency) reported as averages and sentence-level annotations determined by majority vote. We selected two baseline models LLAMA3-8B and LLAMA3.1-70B. To investigate the effects of different approaches, we conducted a comparative analysis of summaries generated by the raw baseline, E2A, Self-Refine and E2A w/ MC-Refine outputs. We additionally include

[4]Due to the high cost of calling GPT-4o APIs (each model evaluation costs more than 5 dollars), we opt for one run and report the significance through paired bootstrap test.

ID	Model	Rel. to A (1-3)	Consis. (1-3)	Asp. Factual. *Fully*	*Part.*	*Not*
1	LLAMA3	2.57	2.56	75.7%	24.3%	*N/A*
2	*w/ Self-Refine*	2.61	2.56	76.1%	23.9%	0.6%
3	E2A	2.60	2.58	77.2%	22.1%	0.7%
4	*w/ MC-Refine*	2.63	2.69	77.6%	22.4%	*N/A*
5	LLAMA3.1-70B	2.52	2.52	73.3%	26.7%	*N/A*
6	*w/ Self-Refine*	2.63	2.56	76.7%	23.3%	*N/A*
7	E2A	2.59	2.63	77.7%	22.3%	*N/A*
8	*w/ MC-Refine*	2.66	2.67	78.5%	21.5%	*N/A*
9	GPT3.5	2.70	2.72	76.6%	22.6%	0.8%
10	GPT4	2.61	2.60	84.5%	15.0%	0.5%

Table 4: Human evaluation results. *Relevance to Aspect* (Rel. to A.) assesses whether the summary discusses the aspect exclusively (3), partially (2), or not at all (1). *Consistency* determines whether the facts in the summary are consistent with the facts in the original input from fully (3) to not supported (1). Additionally, we report the *aspect-based sentence-level factuality (Asp. Factual.)*, which measures the proportion of sentences that are *Fully/Partially/Not* supported by the annotated aspect-focused reflections.

GPT3.5 and GPT4. All systems are anonymized. Annotation details and interface are in Appendix I.

Table 4 shows the performance of different approaches. Consistent with the results in Table 2 and Table 3, different approaches improved the initial summary. The E2A approach obtained performance gains on all three metrics (row 1/5 vs. row 3/7). Unlike the drastic reference-based performance gap between the original and self-refined version, human raters assigned higher relevance scores to self-refined summaries, suggesting that the revisions can help improve the aspect-relevance (rows 2 and 6). Our introduced *E2A w/ MC-Refine* (rows 4 and 8) obtained the best performances on both backbone LLMs, improving both relevance and consistency of the contents, which aligns with the observations from the factuality metrics (Table 2) and GPT-4 based evaluations (Table 3). We observed that sentence-level aspect-based factuality evaluations across different models show similar distributions. This differs from the automatic factuality scores in Table 2, where applying E2A on MC_{EXT} (rows 1-3) improved scores by over 17 points, compared to a 1.5% gain in "Fully Supported" scores for LLAMA3. We attribute this discrepancy to the complexity of sentence-level annotations and the relatively small sample size used in human evaluations compared to the full test set.

5 Analysis of Summary Revisions

GPT-based and human evaluations suggested that revised summaries become more relevant to the

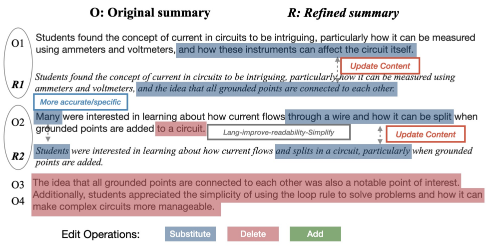

Figure 3: An illustration of edits analysis. Summaries are also in bottom right of Figure 2.

aspect, but how do different refinement approaches help with it? In this section, we present a data-driven study on document-level revisions, aiming to understand what common strategies LLMs use in different refinement approaches. To examine the modifications made by different refinement approaches, we run an automated system (Jiang et al., 2022) to extract edits and determine their underlying intentions. This model (trained on scientific paper revisions) compromises sentence alignment, edit extraction, and intention classification modules. The taxonomy and model details are in Appendix J.1. Figure 3 provides an illustrative example.

We start by exploring the dynamics of **sentence-level edit operations**, aiming to understand how LLMs modify sentences during refinement. As shown in Figure 4, an individual LLM exhibits different behaviors with different refinement strategies. On Llama3.1-70B, DCR favors more proportion of deletions compared to the other approaches (more than 50% of edit operations), which explains its reduced performance on automatic metrics and the shorter content length. For the smaller LLAMA3, *Self-Refine* makes more adding edits compared to the other approaches, potentially introducing details from the original reflections that are not covered by the human reference summaries. When comparing LLAMA3.1-70B and LLAMA3, our proposed E2A w/ MC-Refine approach demonstrates differing proportions of deletion and addition edits, which we attribute to the varying capabilities of LLMs. We include additional analysis on the refinement suggestions and the linguistic features of summaries in Appendices J.2 and J.3.

We also analyzed the **edit intentions** on all revised sentences between original and refined summaries. The distribution of the intentions is visualized in Figure 5. Most edits are categorized as content updates. Notably, DCR exhibits the fewest edits overall (i.e. for LLAMA3.1-70B, DCR contains 856 edits, way less than the other two ap-

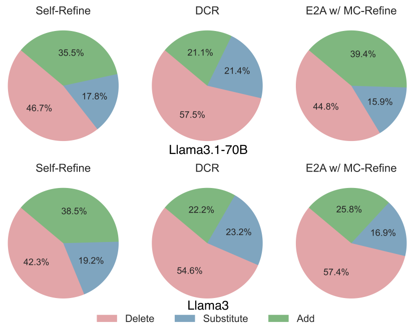

Figure 4: Distribution of edit actions among revised sentences.

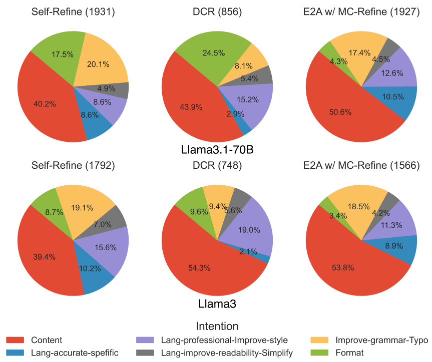

Figure 5: Distribution of span-level edit intentions during different refinements. We include the count of edits in parenthesis.

proaches with around 1900 edits). This matches with previous findings that DCR tends to remove content. Additionally, DCR contains higher proportion of style improvements compared to the other two methods, which does not alter the meaning. Self-Refine instead contains the least proportion of content updates, and includes more grammar typo fixes. The E2A w/ MC-Refine approach has smaller portion of edits for improving the simplicity (4.5% and 4.2%) and improve format (4.3% and 3.4%), which aligns with the goal of enhancing the quality of aspect-based summaries.

6 Related Work

Open Aspect-based Summarization. Recent work introduced multiple OABS datasets (Tan et al., 2020; Amar et al., 2023; Guo and Vosoughi, 2024), where aspects are document-based. Our corpus is the first to have the unique combination of ample human-annotated sentences from the source document with carefully crafted, human-rewritten aspect-based summaries. While proprietary LLM (GPT-3.5) started demonstrating zero-shot capability in performing OABS task (Amar et al., 2023; Guo and Vosoughi, 2024; Mukku et al., 2024), the capabilities of more accessible open-sourced LLMs remains under-explored. Additionally, much of the prior work focuses on domains like news and product reviews, which could potentially be influenced by contamination in the LLMs' training process. *Our study explores the use of open-source LLMs for OABS on a novel dataset featuring diverse, document-dependent aspects in the educational domain, evaluating their performance through comprehensive approaches beyond basic prompting.*.

LLM Feedback and Refinement. Generating feedback at inference-time is essential for LLMs to refine their answers (Madaan et al., 2023; Welleck et al., 2023; Zheng et al., 2023a). More recently, researchers (Huang et al., 2023; Kamoi et al., 2024; Palmeira Ferraz et al., 2024) noticed that LLMs may struggle with self-correction without external guidance. In summarization, prior work (Zhang et al., 2023) leverages GPT-3.5 to iteratively revise summaries to improve the factuality and controllability in news articles. Wadhwa et al. (2024) proposed a specialized "Detect-Critique-Refine" pipeline, which incorporates *fine-tuned* critique and feedback models to enhance the factuality of refined summaries. *We employ open-sourced LLMs to generate feedback based on minimal instructions, leveraging the extractive power of the model and an external fact-checker to better localize errors, to produce better summaries on a given aspect.*

7 Conclusion

In this work, we contribute REFLECTASP, the first open-aspect-based summarization dataset in the educational domain, with 313 high-quality aspect-based summaries and annotated supporting clusters. This dataset offers rich coverage of various aspects, diverse inputs and outputs, and an abstractive nature to enhance human understanding. We extensively test open-sourced LLMs in zero-shot fashions. Our introduced extract-then-abstract (E2A) and refinement approaches improved LLMs' capability in generating more focused summaries, of which results are verified through rigorous automatic and human evaluations. Lastly, our analysis of revisions across different text versions reveals the techniques used by various refinement approaches, offering insights for future innovations in aspect-based summarization.

Limitations

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

Generalizability of the proposed approaches We evaluated our proposed approaches on the SPACE (Angelidis et al., 2021a) dataset, where the MC-Refine approach substantially outperformed other methods (details in Appendix K). However, we identified a bottleneck in the E2A approach, where LLMs are constrained by the context window size when processing long inputs (e.g., over 10k original reviews) and often fail to adhere to the extract-then-abstract instruction. We would like to include the results in the main texts once more spaces are allowed.

Is GPT-based Evaluation Reliable? While there exists a rigorous line of work on prompting GPT models to rate the summaries (Liu et al., 2023; Zhang et al., 2023; Dubois et al., 2023; Wadhwa et al., 2024), we acknowledge that leveraging LLM as an evaluator may carry biases and the model could favor its own output (Zheng et al., 2023b; Panickssery et al., 2024). In our experiments, we use the proprietary LLM, GPT4o as the evaluator to measure the quality of open-sourced models like LLAMA to avoid the potential self-preference. The evaluation prompts are adopted from previous work (Wadhwa et al., 2024).

Manual Annotation Challenges Curating human-annotated datasets presents challenges. As summarized in Table 1, an examination of existing datasets reveals a reliance on automatic collection and alignment methods, such as the heuristic selection of Wikipedia abstracts for OASUM (Yang et al., 2023). This approach often results in incoherent reference summaries. While some work, such as AspectNews (Ahuja et al., 2022) and OpenASP (Amar et al., 2023), repurposed the generic summary through human annotations on aspect labels of sentences, they acknowledged the high cost of such annotations. SPACE (Angelidis et al., 2021a) produced abstractive summarization, proposing a multi-stage pipeline to identify salient sentences and then produce generic/aspect summaries with the help of off-the-shelf classifiers. They reported low inter-annotator agreement. Instead, our dataset came with human-annotated aspect-based clusters with aspect values, which provide more grounding for our collected high-quality human-based abstractive summaries.

Dataset Scale We acknowledge that our dataset only has a test set but no training split. We carefully compared our dataset to many ABS and OABS corpora in §2 Table 1, showing that our dataset is larger than most manually curated corpora for evaluation purpose. Additionally, we want to emphasize that our dataset is derived from real-world scenarios and represents a low-resource setting compared to other domains, such as news and Wikipedia. Collecting annotations from in-house annotators is both time consuming (15 mins per instance) and expensive (suppose the annotator is paid under the minimum wage of $15 per hour, each annotated summary cost around $4). To enhance the future work, we plan to include continuing annotation on the remainder 751 unannotated portion of the 1064 aspect-lecture pairs curated in §2, as well as employing LLMs in synthesizing large scale training data leveraging similar educational datasets.

Utilization of LLM output as the Summary Reference Draft LLMs have been proven to produce high-quality summaries from a human perspective. Goyal et al. (2022) found that summaries crafted by GPT-3 are preferred over those from state-of-the-art (SOTA) fine-tuned models, despite the latter achieving higher scores in reference-based evaluations against human-produced summaries. Furthermore, Pu et al. (2023) found that human evaluators significantly prefer summaries generated by GPT-4, outperforming human-generated summaries and summaries generated by fine-tuned models from multiple perspectives. Our dataset is created as a hybrid of LLM output and human revision, combining the precision of human judgment with the generative capabilities of LLMs to streamline the lengthy process of initial drafting, similar to Liu et al. (2024).

Other Perspectives of Human-Evaluated Text Quality Recent advanced LLMs like GPT-4 are found to produce fluent texts. Besides the mentioned "Relevance to Aspect" annotation, in an older version of the annotation, we prompted the annotators to provide a binary label for fluency, observing that two in-house annotators annotated over 95% of summaries as fluent. This finding aligns with Zhang et al. (2024) and Amar et al. (2023), who found that LLMs generated fluent texts. In our reported human annotation, we introduced both *Consistency* and *Aspect-based Factuality (Sentence-level)* to account for the much more challenging aspect of assessing the factuality of AI-generated summaries. Recent work (Hosking et al., 2024) also demonstrated that human preference scores can under-represent aspects such as factuality, presenting the challenges in better evaluating text qualities in the era of LLMs.

Open-Aspect Extraction for OABS We acknowledge that aspect-based summarization, query-based summarization, and keyword-controlled summarization are intermingled and hard to separate apart fully. While we agree that one can treat the aspect as a keyword that is extracted from the source reflections, open-aspect summarization fits better in this case, as the aspects can be either specific to the source document (i.e., course concepts) or generalizable across different documents such as "Homework" or "Exams." More details on aspect analysis are documented in Appendix B. **Similar to the aspect-based summarization setup in OASUM (Yang et al., 2023), Space and OPPSUM+ (Amplayo et al., 2021), we provided the model with pre-extracted (high-quality human-annotated) aspects in our dataset to guide the generation.** For future work, we would like to add baselines that prompt the model to extract aspects independently before generating aspect-based summaries, though it may introduce another layer of complexity. This approach could lead to more contentious evaluations, given the need to guarantee that differing Large Language Models (LLMs) would need to extract consistent aspects with the human-annotated aspects.

Regarding the **aspect extraction for real world applications**, the ReflectSumm (Zhong et al., 2024) dataset came from a real-world application that has been deployed in real universities. We used their publicly-released dataset to conduct our experiments. The process of aspect extraction, conceptualized as phrase-based summarization, has been addressed in existing literature. Luo et al. have outlined systems for phrase-based summarization comprising candidate phrase extraction, phrase clustering, and phrase ranking (Luo and Litman, 2015; Luo et al., 2016). More recent work has also explored applying LLMs to generate phrase summaries from lecture reflections with promising performances (Zhong et al., 2024). Our research primarily investigates LLMs' capability to produce aspect-based summaries. Thus, we leave the refinement of aspect mining methodologies for future exploration.

Ethical Considerations

Abstractive summarization models have been found to contain hallucinated artifacts that do not faithfully represent the source texts. Regarding the user-sensitive information within the dataset, we do not see concerns about applying our model, as user-specific information will not be included in the students' reflections. The original ReflectSumm dataset was created with students' consent, ensuring their responses could be collected and used for research purposes. We acknowledge the potential for bias in human annotation, particularly in the context of abstractive summaries and crowd-sourced summary evaluations. This is due to the majority of our crowd-sourcing annotators being based in the U.S. However, no private information is collected from the annotators. We only collect annotator's input on the refined summary, as well as evaluations on the qualities. Lastly, the authors acknowledge the use of Grammarly and GPT-4o for correcting sentences that are less fluent but not for generating or drafting new content.

References

2024. *The Role of Reflection-Informed Learning and Instruction in an Introductory Physics Course for Engineering and Science Students*. Zenodo.

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.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021a. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

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

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

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

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations*.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics.

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. *ArXiv*, abs/2310.01798.

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. *Teaching and Teacher Education*.

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. *arXiv preprint arXiv:2310.06825*.

Chao Jiang, Wei Xu, and Samuel Stevens. 2022. arXivEdits: Understanding the human revision process in scientific writing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9420–9435, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, Yuxuan Lai, Chongyang Tao, and Shuai Ma. 2024. Leveraging large language models for NLG evaluation: Advances and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16028–16045, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.

Wencan Luo and Diane Litman. 2015. Summarizing student responses to reflection prompts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Lisbon, Portugal. Association for Computational Linguistics.

Wencan Luo, Fei Liu, and Diane Litman. 2016. An improved phrase-based approach to annotating and summarizing student course responses. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 53–63, Osaka, Japan. The COLING 2016 Organizing Committee.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, 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. *Neurips*.

Muhsin Menekse. 2020. The reflection-informed learning and instruction to improve students' academic success in undergraduate classrooms. *The Journal of Experimental Education*, 88(2):183–199.

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.

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

Sandeep Sricharan Mukku, Abinesh Kanagarajan, Chetan Aggarwal, and Promod Yenigalla. 2024. MARS: Multilingual aspect-centric review summarisation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 894–909, Miami, Florida, US. Association for Computational Linguistics.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA. Association for Computational Linguistics.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *ArXiv*, abs/2309.09558.

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

Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. 2024. Prompt chaining or stepwise prompt? refinement in text summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7551–7558, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. 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 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

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

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal 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.

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.

Manya Wadhwa, Xinyu Zhao, Junyi Jessy Li, and Greg Durrett. 2024. Learning to refine with fine-grained natural language feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12281–12308, Miami, Florida, USA. Association for Computational Linguistics.

Shufan Wang, Laure Thompson, and Mohit Iyyer. 2021. Phrase-BERT: Improved phrase embeddings from BERT with an application to corpus exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10837–10851, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2023. Generating sequences by learning to self-correct. In *The Eleventh International Conference on Learning Representations*.

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 *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023a. Progressive-hint prompting improves reasoning in large language models. *arXiv preprint arXiv:2304.09797*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yang Zhong, Mohamed Elaraby, Diane Litman, Ahmed Ashraf Butt, and Muhsin Menekse. 2024. ReflectSumm: A benchmark for course reflection summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13819–13846, Torino, Italia. ELRA and ICCL.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater.

A Examples on Comparing Prior Work and Our Aspect-based Summaries

In Table 5, we provide more examples in our datasets, where students reflections on "what they find most interesting / confusion" are well summarized in the aspect-based summaries.

B Aspect Analysis

Out of the 1096 phrases collected in *Data Curation*, 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 clusters. 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.

The K-means algorithm we used is from the scikit-learn package (Pedregosa et al., 2011)[5]. 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.

C Human Annotation Details on Reference Summaries

We recruited two in-house annotators (not the authors) to annotate the 313 data points. Both annotators are funded by the project and have taken the courses / possessed the knowledge of course materials covered in the ReflectSumm dataset (Zhong et al., 2024). One of the annotator is a PhD student and the other one is pursing the undergraduate degree.

Two annotation strategies are explored on a first batch of five examples:

(1) Given all the reflections, the annotator needs to first cluster them into different clusters and assign the focused aspect (in noun-phrases). Afterwards, given an assigned aspect, the annotator is tasked to utilize the clusters they built and write a aspect-focused summary from scratch.

(2) Alternatively, the annotators are presented with an AI model generated aspect-summary (here we used the generated version of GPT-4), together with the original aspect. The first step of the annotation is to identify the subset of student reflections that are related to the aspect. Then, the annotators are instructed to check and revise the system summaries to make them aspect-based (focusing on talking about the aspects). The revision also includes removing nonfactual contents that do not exist in the original student reflection, as well as adding contents if they feel are important.

We employed OpenAI's ChatGPT (GPT-4) 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 F.1). The instructions emphasized minimal requirements and explicitly requested the avoidance of unrelated text inclusions.

For the first strategy, the average time spent by the annotators on each instance is 40 min, since assembling and drafting from scratch takes a long time. In contrast, the second strategy took on average 15 min, as the annotators are more focused at identifying the weaknesses of the system summary and focused on producing a high-quality revised version.

D Dataset Analysis Details

D.1 Metric Details

Here we describe the linguistic metrics and would encourage the reader to read the original papers if interested in more technical details.

Content diversity (Grusky et al., 2018) is a joint measure for extractiveness of *coverage* and *density*. Grusky et al. (2018) first proposed an algorithm to compute a set of extractive components between the input and target. *Coverage* measures the per-

[5] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

General Summary (REFLECTSUMM)	**Aspect**	**Aspect-based Summary** (REFLECTASP)
Many of the students today seemed to struggle with the concepts regarding flux and gauss’ law, as well as gaussian surfaces. Some students also struggled with mathematical calculations, while others struggled with the examples that they were doing in class.	Gauss Law/Surface	Students found the topic of Gauss Law and Gaussian surface to be confusing. Although they said that the **lesson on flux and Gaussian surface was tied together seamlessly and easier to understand than their high school lessons,** students still struggle on certain aspects of the topic. Specifically, students **expressed difficulty in choosing a proper Gaussian surface that produces symmetry to allow certain components to cancel out**. Students requested more explanations on the topic to help them understand the topic better.
Most of the responses today included topics about potential and how it confused the students, as well as integrating and setting up their problems that they are given in class. They also had some trouble with some electric field concepts.	Problem Setup	Students struggled with the problem set up in this lecture. On the topic of integrals, students said that solving the integration was understandable and not difficult, **but they were having problems with setting up the integral**. Students **were specifically confused about the limits and variables of integration, and how the limit can change midway through the problem.** Students stated that they could use more clarity and examples of setting up integrals in different situations.
Most students were confused about coming up with designing algorithms and writing pseudo-code for the algorithms. Some students were confused about logistics and content of Milestone 2 and graphing in Matlab. A few students were confused about velocity calculations with regression and the contents of the concept quiz.	Milestone 2	Many students reported feeling confused about the coding aspect of Milestone 2, specifically **regarding which MATLAB functions to use and how to write the pseudo code for the in-class activity**. Some students also expressed confusion about **the noise aspect of the graphs and how to start the coding for the project.** Students were also confused about approaching the project and how to best complete the Milestone, and would have appreciated more clear instructions.

Table 5: Aspect-based summary examples in ReflectASP. In each sample, we **highlight the details** extracted from students’ reflections, which are helpful for the instructor to better assist students’ learning.

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 BST’s"]
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.

centage of words in the summary that come from the source document, while *density* quantifies how well the word sequence of a summary can be described as a series of extractions, given that word orders can be rearranged to construct new contents.

Compression ratio measures the length ratio between the source and target summary, and **Novelty-n** denotes the ratio of new n-grams present in the

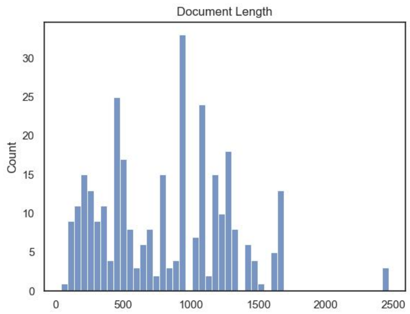

Figure 6: ReflectASP input length distribution.

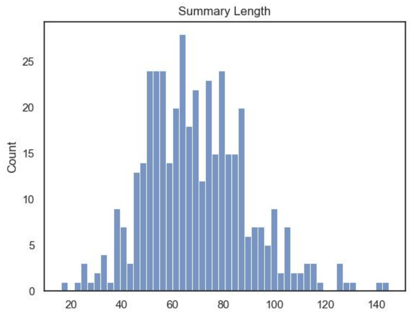

Figure 7: ReflectASP summary length distribution.

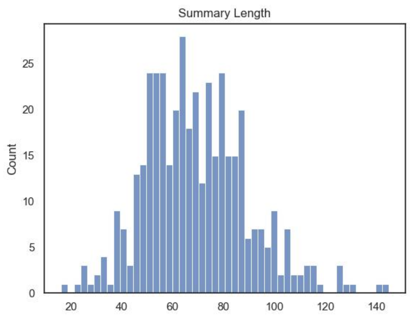

Figure 8: ReflectASP summary's compression ratio.

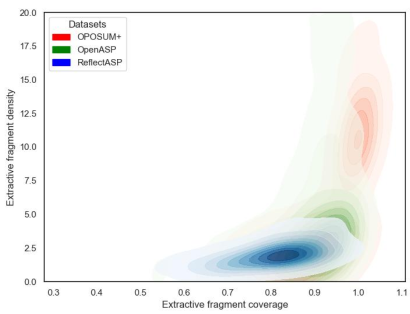

Figure 9: ReflectASP content diversity (Grusky et al., 2018). The area of the plot indicates that, when comparing to the other two corpora, ReflectASP's summary is less extractive (repeating the exact words) but remains faithful based on the high coverage.

summary that are included in the input.

Novelty-n (See et al., 2017) denotes the ratio of new n-grams present in the summary that are not in the input.

D.2 Statistics

Figure 6 and Figure 7 present the distribution of token lengths (measured by word) in input documents and human written summaries. The compression ratio is plotted in Figure 8. Figure 9 compared ReflectASP to the ABS (OPOSUM+) and OABS corpus (OpenASP), showing that ReflectASP contains less direct copying of long spans from the source (lower density) while still retains good enough coverage.

E Model Implementation Details

All of our experiments are conducted on Nvidia L40S GPUs, each with 48 GB RAM. To tackle the memory limitation and speed up the inference with LLMs, we utilize the vLLM (Kwon et al., 2023) to conduct experiments. The Llama3.1-70B needs four cards for inference, and all other models use one card.

E.1 LLMs

We employ LLAMA 2-13B-chat (Touvron et al., 2023)[6], and Mistral-Nemo (Jiang et al., 2023)[7] models for experiments. We additionally include the Llama3-8B-Instruct and Llama3.1-8B-Instruct (`https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`) as well as Llama3.1-70B-Instruct with the quantized version from `neuralmagic/Meta-Llama-3.1-70B-Instruct-quantized.w8a8`. We set the temperature at 0.3 and the max new token at 8000 for generation. We manually evaluated the aspect-based summaries generated during a brief manual tuning of the prompt text to determine the appropriate prompt. We include all model outputs in Appendix F.

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. We additionally included the GPT-4 (gpt4-turbo-0125-preview) and GPT4o (gpt-4o-2024-08-06) models. The temperature is set as 0.5, and the max_token length is set to 256.

[6] `https://huggingface.co/meta-llama/Llama-2-13b-chat-hf`

[7] `https://huggingface.co/nvidia/Mistral-NeMo-12B-Instruct`

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 7: GPT-4 prompt used to generate first draft of reference aspect-based summaries for human revision.

Role	Content
system:	You are a TA for a undergraduate-level course, you are given a collection of student reflections after taking one lecture and tasked to write a summary to present to the instructor
user:	Given the students' responses and a focused topic {aspect}, create a short summary using your own words (no more than 100 words). The summary needs to be a coherent paragraph and should include the major points. The summary should focus on the provided topic only, contain only information about reflections, and avoid adding irrelevant sentences or suggestions such as 'make sure to bring this up in next class', or 'Consider this for future lectures', etc... REFLECTION: {reflections} FOCUSED TOPIC: {aspect} SUMMARY:

Table 8: The baseline prompt used for aspect-based summarization given an aspect.

F Approaches and Prompt Templates

F.1 GPT-4 Prompt for Human-Annotation

We use the prompt in Table 7 to generate system-summaries as the initial draft for human revision.

F.2 Self-Refine Method

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. Our design aligns with the prompt-chaining in (Sun et al., 2024), which was effective and obtained a higher winning rate compared to the step-wise prompt. The prompt for our proposed self-refine framework can be found in Table 9. 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.

F.3 Prompts Used for Experiments

Table 8, 9, 10, and 11 present the final prompts used in our experiments.

F.4 Model Outputs

We include examples of different baseline prompting outputs in Table 12, as well as the comparison of different refinements in Table 13.

G Evaluation Metric Details

ROUGE : We used the implementation in torchmetrics,[8] using stemmer and computing the average when multiple references are available.

BERTScore : We used the implementation from huggingface's evaluate_metrics module[9] and followed the default setup.

MC$_{\text{EXT}}$: We harnessed the SOTA Llama-3.1-Bespoke-MiniCheck-7B (BeSpoke- 1500 MC-7B) released by Bespoke Labs (Tang et al., 2024). The model is fine-tuned from "internlm/internlm2_5-7b-chat" on the combination of 35K data 1503 points following the approach in MiniCheck (Tang et al., 2024). We use the suggested code repo from `https://huggingface.co/bespokelabs/Bespoke-MiniCheck-7B`. Here we paired the human-annotated extractive cluster of aspect-related reflections and a single sentence from the summary as the doc and claim in the fact-checker. We record the predicted labels for that sentence (1 for being factual and 0 for not) and report the macro distribution of labels for all sentences in the 313 generated summmaries.

MC$_{\text{INPUT}}$: This is similar to the setting in MC$_{\text{INPUT}}$, with one exception that we use the full student reflections as the doc for fact-checking.

GPT-4 related metrics : We use the scoring prompt and rubrics from (Wadhwa et al., 2024) and cite their prompts in Figure 10 and Figure 11. We use the GPT4o (gpt-4o-1117 2024-08-06) model.

[8] `https://lightning.ai/docs/torchmetrics/stable/text/rouge_score.html`

[9] `https://huggingface.co/spaces/evaluate-metric/bertscore`

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]
SUGGESTION	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 9: The self-refine prompt with three stages: GENERATION, SUGGESTION, and REFINE.

Stage	Role	Content
	system:	You are a TA for a undergraduate-level course, you are given a collection of student reflections after taking one lecture and tasked to write a summary to present to the instructor
GENERATION	user:	Given the students' responses, and a focused topic {aspect}, create a short summary using your own words (no more than 100 words). The summary needs to be a coherent paragraph and should include the major points. The summary should focus on the provided topic only, contain only information about reflections, and avoid adding irrelevant sentences or suggestions such as 'make sure to bring this up in next class', or 'Consider this for future lectures', etc... You are tasked to perform this task in two steps:(1) Extract the list of indexes and students' reflections in the given REFLECTIONS that are relevant to the focused topic. (2) summarize them into a short summary using your own words (no more than 100 words). REFLECTIONS: {all reflections} FOCUSED TOPIC: {aspect} Your response should be in this json format: {{'Extracted_Reflections' : [your extracted data] (i.e. [STUDENT_REFLECTION_1_TEXT, STUDENT_REFLECTION_2_TEXT, ...]), 'SUMMARY': [your response]}}

Table 10: The E2A prompt used in our task.

Stage	Role	Content
Input		Extracted list of reflections; E2A initial summary
MiniCheck Detect		Run the Minicheck Detector on E2A initial summary sentences, then collect those labeled as not factual in GIVEN_SENTENCES
	system:	You are a TA for a undergraduate-level course, you are given a collection of student reflections after taking one lecture and tasked to write a summary to present to the instructor
SUGGESTION	user:	Given the students' responses, and a focused topic {aspect}, you are provided an extracted list of reflections that are related to the topic and a short summary. (no more than 100 words). ALL REFLECTION: {reflections} FOCUSED TOPIC: {aspect} Extracted List of original reflection and initial Summary: {ext$_{s}ummary$} Initial Summary: {summary} Now, given the below GIVEN_SENTENCES in the summary that is identified as unfaithful, reason if there is any factually inconsistent span in the sentence and propose a way to improve the sentence, making it more concise and focused on the topic {aspect}, utilizing information from the extracted list of reflections. The suggestions should be based on the original reflections and the extracted list of reflections, don't give generic suggestions. ***GIVEN_SENTENCES: {revised_sents}*** ### Task: If GIVE_SENTENCES is []: your response should just return: Suggestions: <no revision needed>. Otherwise, for each sent in GIVEN_SENTENCES, your response should output: Suggestions: original_sent: <SENT1 from GIVEN_SENTENCES>, the error span: <span from SENT1>, the revision suggestion: <your suggested revision>(Delete if there is no need to keep or the post-edit version of SENT1) ### EXAMPLE OUTPUTS: if GIVEN_SENTENCES: [], you should just return "<no suggestion needed>". if GIVEN_SENTENCES= [<SENT1>, <SENT2>], your output would be Suggestions: "{{original_sent: <SENT1>, the error span: <span in SENT1>, the revision suggestion: <Modified version of SENT1>}} {{oiriginal_sent: <SENT2> , the error span: <span in SENT2> ....}}". Reminder: you should only provide Suggestions on sentences from the GIVEN_SENTENCES. If it is empty, just return Suggestions: <no revision needed>.
	sys:	[suggestions]
		Restart the conversation
	system:	You are a TA for a undergraduate-level course, you are given a collection of student reflections after taking one lecture and tasked to write a summary to present to the instructor
REFINE	user:	Given the students' responses, and a focused topic {aspect}, a list of extracted responses that are relevant to the topic, and some suggestions on revisions, you are tasked to improve a short summary. (no more than 100 words). Please improve the short summary written below, incorporating the suggestions. The suggestions are on sentences of the summary, so please only modify those highlighted sentences and keep the remainder unchanged. The revised version should be no more than 100 words, focusing on the topic of {aspect} based on below reflections, The summary should be a coherent paragraph and should include the major points. If you think the initial summary is good enough, you can make minimal changes. You also need to pay attention to the extracted list of reflections that are related to the given topic {aspect}: ALL REFLECTION: {reflections} FOCUSED TOPIC: {aspect} Extracted List of original reflection and Initial Summary: {summary} Revision Suggestions: {suggestions} You should pay attention to the revision suggestions and decide whether you want to edit the sentences with suggested revisions. If the Revision Suggestions mentions "no suggestion needed", you should not revise the initial summary. Your summary should have minimal changes on the initial summary and be factual. Your response should just output the refined summary and should not include any extra explanation on changes. The format is – REFINED SUMMARY: [your response]

Table 11: The E2A w/ MC-Refine prompt used for open-aspect-based summarization (OABS).

Model	Summary
LLAMA2	The students found the light bulb demonstration and the plastic sheet experiment to be the most interesting and helpful in understanding the concepts of electric charge and forces. They appreciated the visual aids and the interactive nature of the lecture, including the clicker questions and breakout room discussions. Many students found the similarities between electric and gravitational forces to be fascinating, and the concept of electric charge properties was well received. The demonstration of energy conservation through the splitting of a gamma ray photon into an electron and positron was also found to be intriguing. Overall, the students enjoyed the interactive and visual aspects of the lecture and found them to be effective in deepening their understanding of the topics covered.
LLAMA3	The students found the lightbulb demonstration to be a highlight of the lecture, with many mentioning that it helped them understand the concept of electric forces and charges. The visual representation of the demonstration made it easier for students to grasp the abstract concepts, and some even found it "fun to watch". The similarity between electric and gravitational forces was also a point of interest, with students noting that it helped them understand the concept better. Additionally, the interactive nature of the lecture, including the use of breakout rooms and clicker questions, was appreciated by many students. Overall, the lightbulb demonstration and related discussions seemed to have a significant impact on students' understanding of electric charges and forces.
LLAMA3.1	Students found the light bulb demonstrations to be engaging and effective in illustrating complex concepts, such as the separation of charges and the distribution of electric force. Many appreciated the visual representation and hands-on experience, which helped to clarify their understanding of these abstract topics. Additionally, the demonstrations sparked interesting comparisons between electric and gravitational forces, highlighting the similarities between the two. Overall, the interactive nature of the lecture, including breakout rooms and practice problems, was well-received and contributed to a deeper understanding of the material.
LLAMA3.1-70B	Students found the lightbulb demonstration to be a highlight of the lecture, with many noting that it provided a clear visual representation of electrostatic concepts. The separation of charges and the resulting lighting of the bulb was particularly interesting to several students. The demonstration was also seen as engaging and fun to watch, helping to capture students' attention and make the topic more accessible. Overall, the use of visual aids and hands-on demonstrations was well-received by students, who appreciated the ability to see complex concepts in action.
MISTRAL	Students found the light bulb demonstrations engaging and helpful in understanding abstract concepts like electric force and charge separation. They appreciated the visual representations, with several mentioning that it made complex topics easier to grasp. The interactive nature of these demonstrations, allowing them to see the results of changing variables, was particularly appealing. Some students, however, noted that the pace of explanations during these demonstrations could be overwhelming, suggesting a need for slower, clearer communication.
Human Reference	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.

Table 12: An example of different baseline system summaries. The aspect is “Light Bulb/Demonstration” .

Model	Summary
LLAMA 3.1-70B	Students found the lightbulb demonstration to be a highlight of the lecture, with many noting that it provided a clear visual representation of electrostatic concepts. The separation of charges and the resulting lighting of the bulb was particularly interesting to several students. The demonstration was also seen as engaging and fun to watch, helping to capture students' attention and make the topic more accessible. Overall, the use of visual aids and hands-on demonstrations was well-received by students, who appreciated the ability to see complex concepts in action.
LLAMA 3.1-70B w/ Self-Refine	Students found the lightbulb demonstration to be a highlight of the lecture, providing a clear visual representation of electrostatic concepts, particularly the separation of charges and the resulting lighting of the bulb. Many students noted that the demonstration was engaging and fun to watch, making complex concepts more tangible. The visual representation of electrostatic concepts, such as the separation of charges, was especially interesting to several students, helping to illustrate key principles in a memorable way.
LLAMA 3.1-70B w/ DCR	Students appreciated the use of visual aids and hands-on demonstrations in the lecture. The light-bulb demonstration was a highlight of the lecture, helping to capture students' attention and make the topic more accessible. Overall, the use of visual aids and hands-on demonstrations was well-received by students, who appreciated the ability to see complex concepts in action.
LLAMA 3.1-70B E2A	Students found the lightbulb demonstrations to be engaging and interesting, with the visual representation helping to clarify complex concepts. The separation of charges and the experiment setup caught their attention, making the topic more enjoyable and easier to understand. The demonstrations were considered a valuable learning tool, providing a clear and captivating way to illustrate key principles.
LLAMA 3.1-70B E2A w/ Self-Refine	Students found the lightbulb demonstrations to be engaging and interesting, with the separation of charges and experiment setup catching their attention and making the topic more enjoyable and easier to understand. The demonstrations provided a valuable learning tool, offering a clear and captivating way to illustrate the effects of electric forces in action, making complex concepts more tangible and relatable.'
LLAMA 3.1-70B E2A w/ MC-Refine	Students found the lightbulb demonstrations to be engaging and interesting, as they provided a clear visual representation of the concept. The separation of charges and the experiment setup caught their attention, making the topic more enjoyable and easier to understand. The demonstrations were considered a valuable learning tool, providing a clear and captivating way to illustrate key principles.
Human Reference	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.

Table 13: An example of different refined summaries. The aspect is "Light Bulb/Demonstration" .

Prompt D.1: Zero-Shot prompt used with GPT-4 to generate a factuality score on a scale of 1-5.

Determine whether the provided summary is consistent with the corresponding document. Consistency in this context implies that all information presented in the response is substantiated by the document. If not, it should be considered inconsistent.
{{ instruction }}
{{ response }}
The response can have one or more of the following errors:
1. Extrinsic Information: the response contains new information not grounded in the source material
2. Mis-Referencing: a property or an event in the response can be found in the source material, but are associated with the wrong entity
3. Stating Opinion As Fact: the response entails a proposition that's mentioned in the source material not as a fact, but as someone's opinion
4. Reasoning Error: the response makes one or more wrong inferences from the information in the source material
5. Tense/modality Error: the tense or modal (eg: can, may, must) used in the response sentence does not match the tense/modality of the source material
6. Contradiction: the response contradicts the source material
7. Nuanced Meaning Shift: the response twists information from the source material in a subtle way
Given the error categories, rate the above response on a scale of 1 to 5 based on extent of factual consistency:

5. completely consistent: the response is completely factually consistent with the source material.
4. insignificant inconsistencies: the response is mostly factually consistent, with slight inconsistencies not affecting main points.
3. partially inconsistent: overall factually consistent, with a few inconsistencies with the source material.
2. severe inconsistencies: nearly half response is factually inconsistent, with severe deviation from main points.
1. completely inconsistent: the entire response is factually inconsistent with the source material.
First output a list of errors that the summary makes, then conclude the response with a score in the following format: "therefore, the score is:"

Figure 10: Zero-shot prompt used with GPT-4 to generate single factuality scores on a scale of 1-5. The figure is created by Wadhwa et al. (2024)

Prompt D.2: Zero-Shot prompt used with GPT-4 to generate pairwise factuality scores on a scale of 1-5.

Determine whether the provided summary is consistent with the corresponding document. Consistency in this context implies that all information presented in the response is substantiated by the document. If not, it should be considered inconsistent.
{{ instruction }}
Response 1 {{ response1 }}

Response 2 {{ response2 }}

A response can have one or more of the following errors:
1. Extrinsic Information: the response contains new information not grounded in the source material
2. Mis-Referencing: a property or an event in the response can be found in the source material, but are associated with the wrong entity
3. Stating Opinion As Fact: the response entails a proposition that's mentioned in the source material not as a fact, but as someone's opinion
4. Reasoning Error: the response makes one or more wrong inferences from the information in the source material
5. Tense/modality Error: the tense or modal (eg: can, may, must) used in the response sentence does not match the tense/modality of the source material
6. Contradiction: the response contradicts the source material
7. Nuanced Meaning Shift: the response twists information from the source material in a subtle way

Given the error categories, rate each response on a scale of 1 to 5 based on extent of factual consistency:
5. completely consistent: the response is completely factually consistent with the source material.
4. insignificant inconsistencies: the response is mostly factually consistent, with slight inconsistencies not affecting main points.
3. partially inconsistent: overall factually consistent, with a few inconsistencies with the source material.
2. severe inconsistencies: nearly half response is factually inconsistent, with severe deviation from main points.
1. completely inconsistent: the entire response is factually inconsistent with the source material
For each response, first output a list of errors that the summary makes, then conclude the response with a score in the following format: "therefore, the score is:"

Output Format:

Response 1
...

Response 2
...

Figure 11: Zero-shot prompt used with GPT-4 to generate pairwise factuality scores on a scale of 1-5. The figure is created by Wadhwa et al. (2024).

H More Automatic Results

Due to space limit, we omit one variant **E2A w/ Self-Refine**, that applies the *Self-Refine* approach on E2A results in Table 2 and two weaker LLM baselines. Here we included them with some discussions.

H.1 Additional Reference-based Evaluation Results

We include extra backbone models (LLAMA2 and MISTRAL, as well as the additional **E2A w/ Self-Refine** rows in the extended version of Table 2.

Older models generate worse summaries evaluated by ROUGE and BERTScore, meanwhile, the E2A approach does not help as the models fail to follow instructions in generating extracted list. Additionally, as seen in Table 14, applying Self-Refine on E2A outputs (rows 9-11) experienced performance drops on ROUGE and BERTScores when compared to the original version (rows 4, 10 and 16 in Table 2).Similar results were observed in factuality scores, with one exception in LLAMA3.1-70B, where the MC scores improved. This indicates that stronger models can generate more useful suggestions, leading to more effective revisions.

H.2 Additional Automatic GPT-based Evaluation

Table 15 shows the additional GPT-based automatic evaluation results. Smaller models may suffer from generating non-helpful suggestions, which leads to a drop of summarization quality (LLAMA3 E2A vs. w/self-refine in the first chunk of the bottom block).

ID	Model	R-1	R-2	R-L	BS	MC_{EXT}	MC_{INPUT}	# Sents	# Words
1	LLAMA2	**45.05**	**20.61**	**41.05**	89.61	34.13	80.93	5.63	130.4
2	*w/ Self-Refine*	36.04	14.68	32.75	88.20	26.07	59.31	7.42	185.7
3	*w/ DCR*	42.32	16.46	38.27	**89.76**	**55.33**	**86.89**	3.62	59.8
4	**E2A**	40.55	16.44	36.20	88.62	28.47	62.82	7.55	185.2
5	MISTRAL	**43.55**	13.66	**38.40**	**89.50**	44.33	**85.10**	3.67	78.71
6	*w/ Self-Refine*	40.05	11.33	34.98	88.81	39.88	74.01	3.43	83.4
7	*w/ DCR*	32.57	10.17	29.28	89.25	61.22	86.75	2.35	34.8
8	**E2A**	43.26	**14.55***	38.06	89.45	**53.25***	71.91	3.62	63.2
9	LLAMA3 **E2A** *w/ Self-Refine*	44.32	15.68	38.78	89.63	49.43*	85.02*	2.95	79.3
10	LLAMA3.1 **E2A** *w/ Self-Refine*	43.68	15.29	38.41	89.53	38.98	79.10	3.18	90.1
11	LLAMA3.1-70B **E2A** *w/ Self-Refine*	44.71	16.14	39.36	89.69	48.26*	80.99	3.21	87.9

Table 14: Extra experimental results on REFLECTASP: BS refers to BERTScore F1. MiniCheck scores are reported. For ROUGE (R-1/2/L), BS, and MC scores, all results are averaged over three runs, and * means the score is significantly better than the baseline models within each block. Gray rows indicate the baseline models, The best and second best scores for each backbone are **bold** and underlined.

Model	G↑	ΔG↑	Pair ΔG↑	*W*↑	*S*	*L*
Pairwise Comparison with Baseline Summary as the Original Input						
MISTRAL	2.74	-	-	-	-	-
w/ Self-Refine	2.62	-0.14	-0.18	0.12	0.65	0.23
E2A	2.84†	0.08*	0.20*	0.27	0.64	0.09
LLAMA3	2.74	-	-	-	-	-
E2A *w/ Self-Refine*	2.80	0.03*	0.12*	0.23	0.64	0.13
LLAMA3.1	2.76	-	-	-	-	-
w/ Self-Refine	2.67	-0.09	0.03*	0.18	0.67	0.15
w/ DCR	2.80	0.01	**0.33***	**0.45**	0.42	0.13
E2A	2.88†	0.07*	0.22*	0.27	0.66	0.07
E2A *w/ MC-Refine*	**2.89**	**0.10***	0.18*	0.24	0.69	0.07
LLAMA3.1-70B	2.85	-	-	-	-	-
E2A *w/ Self-Refine*	2.88	0.02	0.14*	0.24	0.66	0.10
Pairwise Comparison with E2A Summary as the Original Input						
LLAMA3 E2A	2.85	-	-	-	-	-
w/ Self-Refine	2.81	-0.03	-0.22	0.06	0.68	0.26
LLAMA3.1 E2A	2.87	-	-	-	-	-
w/ MC-Refine	2.88	0.03	-0.01	0.07	0.84	0.07
LLAMA3.1-70B E2A	2.87	-	-	-	-	-
w/ Self-Refine	2.88	0.00	0.00	0.12	0.78	0.10

Table 15: Additional GPT-related evaluation results of different methods. Within each block (added to Table 3, pairwise metrics compare the outputs of the given system to those in the highlighted rows. A † indicates significant improvement over the previous row (p<0.05) based on a paired bootstrap test, while * denotes that the absolute value is significantly different from zero. We see that our proposed *E2A w/ MC-Refine* achieves the largest / second largest gains across all metrics when compared to the baseline, and it is significantly better than the original E2A-generated summary.

H.3 How does E2A MC-Refine benefit

In this section, we examine the effectiveness of the quality of extracted supporting reflections on model performance.

Model	R-1	R-2	R-L
LLAMA3	61.25	51.83	60.59
LLAMA3.1	61.16	54.19	60.66
LLAMA3.1-70B	79.78	75.26	79.57
Human	79.06	73.52	78.68

Table 16: Performance of different LLMs' extracted supporting reflections using the E2A approach comapred to human extracted clusters. ROUGE (Lin, 2004) is used an a proxy to inter-annotator agreement.

We evaluate the quality of different LLMs' extracted reflections against the human annotated ones. We estimate the human upper bound by measuring the ROUGE score between double-annotated clusters and report the ROUGE score, following Angelidis and Lapata (2018). As shown in Table 16, LLAM3.1-70B extracted the set of aspeect-related reflections with near human performance.

I Amazon Mechanical Turk Crowd-sourcing Evaluation Details

We source crowd-workers from Amazon Mechanical Turk, requiring them with more than 95% approval rate and more than 5000 approved HITS. Workers were instructed to thoroughly read the annotation guidelines, which included examples and are illustrated in Figures 12 to 14. We ensured that our compensation met the minimum hourly wage requirement (currently anonymized for reviewing purposes). We released five test samples which are know to contain non-factual errors (each can be annotated for at most 20 times) to a pool of 100 workers who have been listed on a white list as the qualification task. We filtered unqualified workers who selected score of 3 for the summary-level evaluation criteria or did not participate.

We have in total of 1500 annotations, spanning 500 summaries and 24 annotators, which resulted in an extreme sparse annotation matrix. Meanwhile, over 1100 of each document-level label (Relevance to Aspect and Consistency) are dominated by the value of 3, making the annotation task highly imbalanced. This makes reliability testing more challenging, with mainstream inter-annotator agreement scores, as measured by Krippendorff's Alpha (Krippendorff, 2011), nearing 0.1. One approximation for the agreement is perfect-agreement (which means all 3 annotators picked the same score). For *Relevance*, 97 summaries are rated with all same scores, and the remainder 247 have a majority voting of 3. For *Consistency*, 101 have the perfect-agreement and 246 has majority label of 3, suggesting that reviewers can have high agreement in picking the score of 3. While prior work (Angelidis et al., 2021a; Zhang et al., 2023; Amar et al., 2023) did not report the quality of human annotations, we investigated the challenges of measuring large-scale sparse annotations and emphasized the need for better-designed human evaluation protocols for future work.

The Amazon Mechanical Turk annotation interface is in Fig 15. The full annotation cost over 200 US dollars.

J Supplementary Materials about Revision Analysis

J.1 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 specificness 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 a large amount of scientific content, add or delete major facts.

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

Introduction to the Human Evaluation Study

Your task involves assessing the quality of a summary 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
   a. the **Relevance to the aspect:**

      Determine if the summaries address the specified aspect exclusively, partially, **or not at all.**

   b. The **Overall Consistency**

      Determine if the facts in the summary are consistent with the facts in the original article. You should pick from **fully supported (3), partially supported (2)** or **not supported (1)**.

   c. **Sentence-level Aspect-based Factuality:**

      Look at the provided highlighted input reflections **(which were manually selected as reflections that are focused on the given aspect)**.

      For **each sentence in the summary,** determine whether the contents mentioned in the summaries are indeed discussed in the reflections. *You should pick from* **fully supported (3), partially supported (2)**, or **not supported at all (1)**.

Examples

Input Reflections

Aspect: Gaussian Surface

(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?

Figure 12: A screenshot of the human annotation guideline to evaluate system generated aspect-based summaries. (1/3)

Example 1: (Relevance to the aspect)

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

You do not need to verify whether the content covering the aspect is factual or consistent with the source texts; this will be addressed in the third question.

Aspect: Gaussian Surface

Summary 1.
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.**

Overall Consistency: Partially Support (2)
(Rationale): ***The red italic sentence*** can not find supporting fact from the full student reflections, thus, the summary is partially supported.

Aspect: Gaussian Surface

Generated Summary 2:
Students had difficulty understanding the concept of Gaussian surface, particularly when it comes to choosing the shape of the surface and how it relates to the orientation of electric fields. They also found the problem of figuring out the flux through a Gaussian surface challenging, and were unsure about the different symmetries involved. Some students were confused about the concept of conductor vs insulator in reference to Gauss' law, while others found the orientation of planar electric fields and cylindrical Gaussian surface to be confusing.

Relevance to the aspect: Exclusively
(Rationale): Each sentence is discussing about students' reflections on the topic of Gaussian Sur and related topics.
Overall Consistency: Fully Supported

Figure 13: A screenshot of the human annotation guideline to evaluate system generated aspect-based summaries. (2/3)

Example 2: (Aspect-based Factuality)

Look at the **highlighted input reflections. For each sentence in the summary,** determine whether the contents mentioned in the summaries are indeed discussed in the reflections. *You should pick from* **fully supported (3), partially supported (2)**, or **not supported at all (1).**

Original Student Reflections

Aspect: Gaussian Surface

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

- 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
- Figuring out what Gaussian surface to draw in problems
- 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.
- Are these the only three symmetries we will work with?

Aspect: Gaussian Surface

Aspect-based Factuality:

[1] The students found the concept of Gaussian surface to be challenging, particularly the choice of the surface and its relationship to the electric field.
3: Fully Supported
[2] They also struggled with the idea of flux and how it is affected by positive and negative charges being inside or outside the surface.
1: Not Supported at all
[3] Additionally, the students found the orientation of planar electric fields and cylindrical Gaussian surface to be confusing.
3: Fully Supported
[4] 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.
2: Partially Supported
[5] They also recommended that the instructor provide more clarification on the choice of the surface and how it affects the electric field.
3: Fully Supported
[6] 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.
2: Partially Supported

(Rationale):
There are in total of 6 sentences.

Fully supported **(green)**: sentence 1, 3, and 5
Partially Supported **(yellow)**: sentence 4, 6
not supported at all **(purple)**: sentence 2

Sentences 1, 3, and 5 can be connected to the highlighted reflections and thus are **fully supported.**
Sentence 4 points out the understanding of the concept of Gaussian surfaces and their application to electric fields, which is **partially connected** to the student's reflection: The only part I found a little difficult was the different gauss fields around the uniformly charged inner sphere ... ; In detail, The first part of [4] sentence with **more examples and practice problems** is not mentioned in the highlighted reflections.
Sentence 6 is **partially connected**, "the topic of Gassian surface to be confusing" can be supported by the highlighted sentence, yet he inquiry of "more practice and clarification , they were able to understand basic concept" did not exist in the students reflections.
Sentences 2 **(highlighted in purple)**does not have any supporting sources in **the highlighted student reflection.**

Figure 14: A screenshot of the human annotation guideline to evaluate system generated aspect-based summaries. (3/3)

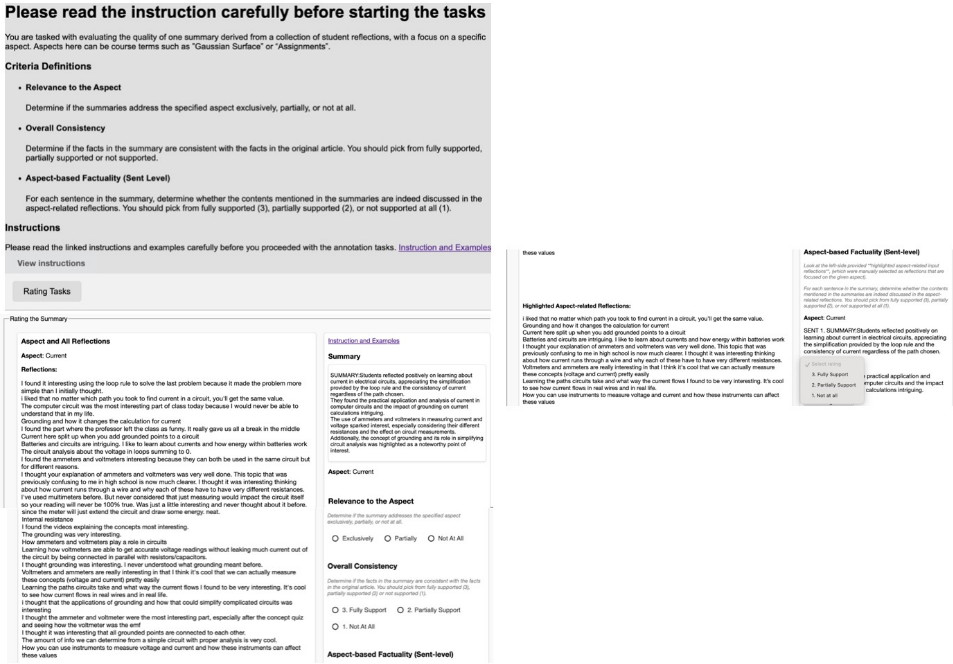

Figure 15: A screenshot of MTurk HIT used to evaluate the quality of system generated aspect-based summaries.

We recommend the reader check 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.

J.2 Suggestion Analysis

Besides directly analyzing the difference between versions of summaries, it is also crucial to inspect the suggestions provided by different refinement approaches. In Table 17, we present samples of suggestions for three systems, Self-Refine, DCR, and MC-Refine on LLAMA3 and LLAMA3.1-70B outputs.

We find that different refinement approaches have distinct revision suggestions. For Self-Refine, 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’, supported with model generated concrete revision demonstrations. DCR suggestions are also in the list format. Given that they are trained on a synthesized datasets, these suggestions follow the format of (1) evaluating the sentence in the context of full summary, (2) identifying the error span and (3) suggesting changes of a single sentence. Examples in the following paragraph demonstrated that DCR’s suggestions may focus on removing less confident details, which may not help improve the summarization quality. Lastly, the MC-Refine approach’s suggestion comes with a format: original_sent, error_span, and the revision suggestion. We also notice that stronger LLMs can provide reasoning on the revisions suggested.

In Table 17, we present samples of suggestions for three systems, Self-Refine, DCR, and MC-Refine on LLAMA3 and LLAMA3.1-70B outputs. Adhering to the findings in §5, we find that DCR’s suggestions focused on removing details from the original summary (highlighted in blue colors). For instance, in the second point of column 1 row 2, DCR’s suggestions is *The error span is: ’They appreciated the demonstration of how current flows through a circuit, including the role of grounding and how it affects the calculation.’ To fix this, consider changing the span to ’Students found the concept of current to be fascinating’*. Such removal, while improving the factuality metrics, did not help improve the quality of the aspect-based summaries. In contrast, *Self-Refine* employed diverse strategies to improve the specificity, removing non-related

Method	LLAMA3	LLAMA3.1-70B
Self-Refine	Suggestions: **1. Focus on the most prominent points**: While the original summary covers a wide range of topics, it could be condensed to focus on the most prominent points. For example, the loop rule, grounding, and the role of ammeters and voltmeters could be highlighted as the main takeaways. **2. Remove repetitive language:** Some phrases, such as "I found it interesting" or "I thought it was interesting," are repeated throughout the reflections. These could be condensed or removed to make the summary more concise. **3. Emphasize the key concept**: The topic of current is the main focus of the reflections, but the summary could be more explicit in highlighting this concept. For example, the summary could start with a sentence like "The concept of current was a central theme in today's lecture, with students exploring its calculation, measurement, and applications.	Suggestions: 1. To make the summary more concise, consider removing the phrase "Additionally, students appreciated the simplicity of using the loop rule to solve problems and how it can make complex circuits more manageable" as it slightly deviates from the focused topic of Current and is not as directly relevant as the other points mentioned. 2. To improve focus on the topic of Current, rephrase the sentence "Many were interested in learning about how current flows through a wire and how it can be split when grounded points are added to a circuit" to "Students were interested in learning about how current flows and splits in a circuit, particularly when grounded points are added" to make it more direct and concise. 3. To make the summary more concise, consider combining the ideas of the first and third sentences into one sentence, such as "Students found the concept of current in circuits to be intriguing, particularly how it can be measured using ammeters and voltmeters, and the idea that all grounded points are connected to each other."
DCR	1. The summary introduces specific details about the loop rule and its application that are not present in the given reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'the idea that the loop rule simplifies the calculation of current in a circuit' To fix this, consider changing the span to 'Students found the concept of current to be fascinating' 2. The summary introduces specific details about the demonstration of current flow and the role of grounding that are not mentioned in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'They appreciated the demonstration of how current flows through a circuit, including the role of grounding and how it affects the calculation.' To fix this, consider changing the span to 'Students found the concept of current to be fascinating.' 3. The summary introduces specific details about the use of ammeters and voltmeters that are not mentioned in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'the use of ammeters and voltmeters' To fix this, consider changing the span to 'the importance of measuring voltage and current' 4. The summary introduces information about the internal resistance of batteries and its effect on current flow that is not present in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is:'students found it interesting to learn about the internal resistance of batteries and how it affects the flow of current' To fix this, consider changing the span to 'students appreciated the demonstration of how current flows through a circuit, including the role of grounding and how it affects the calculation' 5. The summary introduces information about students' enjoyment and exploration of the concept of current, which is not present in the provided reflections. This information is extrinsic and adds new details not grounded in the document. The error span is:'students enjoyed exploring the concept of current and its applications in real-life circuits.' To fix this, consider changing the span to 'Students found the concept of current to be fascinating.'	1. The summary introduces specific details about the measurement of current using ammeters and voltmeters and how these instruments affect the circuit, which are not mentioned in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'Students found the concept of current in circuits to be intriguing, particularly how it can be measured using ammeters and voltmeters, and how these instruments can affect the circuit itself.' To fix this, consider changing the span to 'Students found the concept of current in circuits to be intriguing.' 2. The summary introduces specific details about how current flows through a wire and how it can be split when grounded points are added to a circuit, which are not mentioned in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'Many were interested in learning about how current flows through a wire and how it can be split when grounded points are added to a circuit.' To fix this, consider changing the span to 'Students found the concept of current in circuits to be intriguing.' 3. The summary introduces specific details about the concept of current in circuits that are not directly mentioned in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'the concept of current in circuits to be intriguing, particularly how it can be measured using ammeters and voltmeters, and how these instruments can affect the circuit itself' To fix this, consider changing the span to 'Students found the concept of current in circuits to be intriguing' 4. The summary introduces information about the loop rule that is not present in the provided reflections. This information is extrinsic and adds details not grounded in the document. The error span is: 'the simplicity of using the loop rule to solve problems and how it can make complex circuits more manageable' To fix this, consider changing the span to 'the simplicity of using the loop rule to solve problems'
E2A w/ MC-Refine*	Suggestions: {**original_sent:** Students found the concept of current to be interesting and engaging, particularly in the context of circuits and the role of ammeters and voltmeters in measuring voltage and current. They appreciated the idea that current remains constant regardless of the path taken to find it, and were fascinated by the ways in which grounded points can affect the flow of current. Additionally, students were intrigued by the impact of measurement instruments on the circuit itself and the importance of considering this when taking readings., **the error span:** the last part of the sentence, **the revision suggestion:** The students were also intrigued by the practical applications of current in real-life circuits and the role of instruments in measuring it accurately.}	Suggestions: {**original_sent**: Students found the concept of current in circuits fascinating, particularly how it flows through wires, splits when grounded points are added, and can be measured using ammeters and voltmeters., **the error span**: how it flows through wires, **the revision suggestion**: Students found the concept of current in circuits fascinating, particularly how current splits when grounded points are added, and how it can be measured using ammeters and voltmeters. Reasoning: The original sentence mentions "how it flows through wires" which is not explicitly mentioned in the extracted list of reflections. While it is a related concept, the reflections focus more on the measurement and behavior of current in circuits. The revised sentence focuses on the aspects of current that are directly mentioned in the reflections, such as splitting when grounded points are added and measurement using ammeters and voltmeters.}

Table 17: Suggestions generated by different refinement approaches using two backbones LLAMA3 and LLAMA3.1-70B. The examples remained the same to Fig 2. * means that the original version for the E2A w/ MC-Refine is E2A generated summaries, while the former two are generated by the baseline.

phrases, as well fusing contents to make it more concise. As for *E2A w/ MC-Refine*, leveraging high-quality extracted supporting reflections, and the fact-checker that help localize the errors, this approach made minimal edits and retain the contents that are deemed good quality.

J.3 Linguistic Analysis on Different Model's Summaries

Figure 16 presents a linguistic analysis of system diversity in relation to the original reflections. For compression ratio (second column), DCR demonstrates the most significant compression, consistent with the findings in §4.1. Notably, system-generated aspect-based summaries exhibit higher Novelty-N values compared to human-edited summaries (0.19/0.63/0.83 for Novelty-1/2/3, respectively), indicating that these systems incorporate vocabulary not present in the original reflections.

Additionally, the E2A approach expands the area of the plots, suggesting that the generated summaries have more diverse content compared to the baseline. This improvement can be attributed to the effective extraction of supporting reflections during the extraction phase.

A pairwise comparison between E2A and E2A w/ MC-Refine reveals that larger models tend to produce smaller plot areas, reflecting increased text diversity, which consequently enhances summary quality.

K Other Domain Results

In addition to our REFLECTASP dataset, we additionally tested the proposed approaches on a out-of-domain dataset, SPACE (Angelidis et al., 2021a). We used their testing split, which contains reviews about 25 unique hotels. Each set contains 100 real reviews, and aspect-based summaries on each of six popular aspects: *building*, *cleanliness*, *food*, *location*, *rooms*, and *service*.

K.1 Baselines

We used the baseline system outputs released by (Hosking et al., 2023), including **QT** $_{\text{asp}}$(Angelidis et al., 2021b), **HERCULES**$_{\textbf{ext}}$ and **HERCULES**$_{\textbf{abs}}$ (Hosking et al., 2023), as well as **AceSum**$_{\textbf{ext}}$ and **AceSum** (Amplayo et al., 2021).

We include the following methods: (1) Baseline, (2) Self-Refine, and (3) MC-Refine applied to the baseline outputs. Initial results show that LLMs struggled to process longer inputs (an average of 14k words for the combined length of all 100 reviews) and to follow instructions for generating E2A results. For the MC-Refine detector, we utilize the ground-truth clusters provided in the dataset.

K.2 Evaluation

We report the same set of evaluation metrics as in §4.1. One difference is that we follow the ROUGE implementation in Hosking et al. (2023), which is the *'jackknifing'* method for multiple references as implemented for the GEM benchmark (Gehrmann et al., 2021), to make the evaluation consistent with scores reported in prior papers.

K.3 Result

Table 18 showed the results of different approaches. While the older baselines (rows 1–5) achieved higher ROUGE and BERTScore metrics, their factuality scores were lower than those of LLMs. We attribute this discrepancy to the lexical differences between the older supervised models and the zero-shot LLMs. As shown in Table 19, the outputs of the best baseline model, AceSum(abs), closely resemble human references, while LLM outputs tend to include more details from the original reviews.

Our proposed *MC-Refine*, combined with human-selected clusters for fact-checking, obtained substantial performance gains compared to the baseline and other methods. These improvements were more notable on the larger LLAMA3.1-70B model (row 16 vs. row 13). Meanwhile, the factuality of generated contents is also improved.

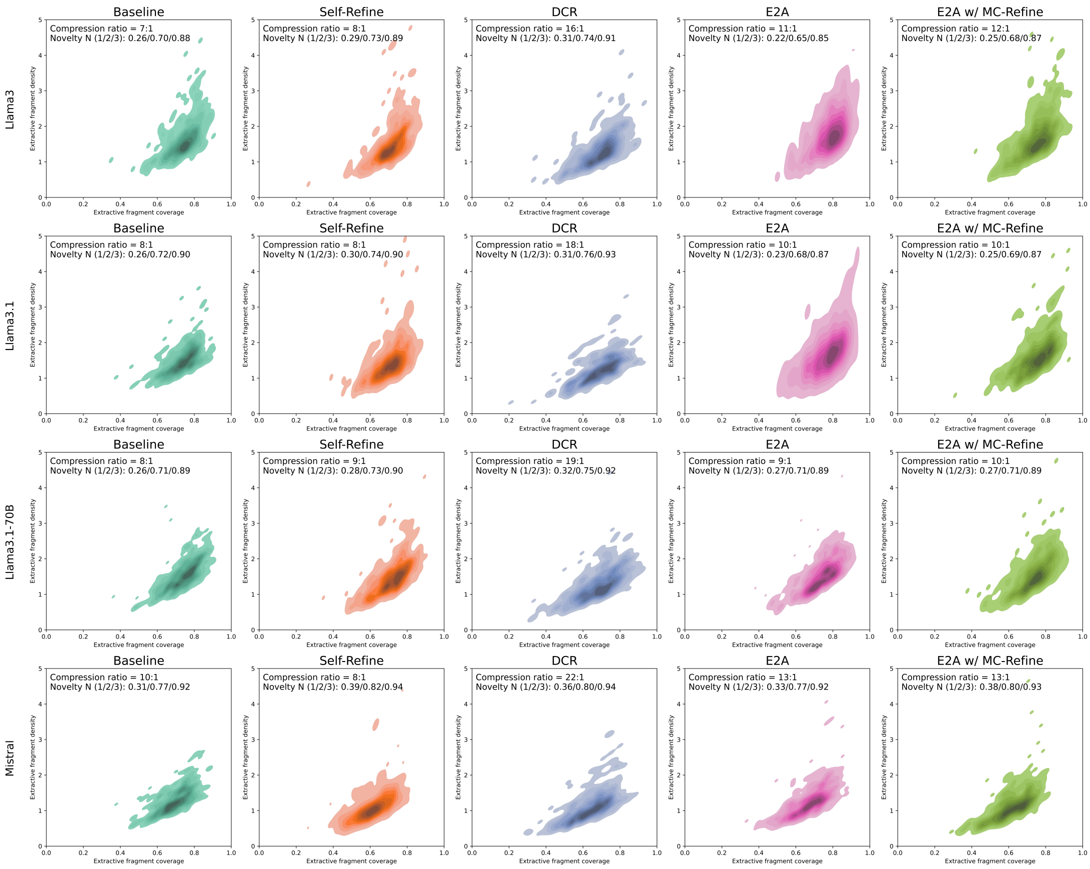

Figure 16: Density and coverage distributions of generated summaries across different models and LLM backbones. We additionally report the compression ratio and novelty-n metrics on the top-left corner. Each box is a normalized bivariate density plot of extractive fragment coverage (x-axis) and density (y-axis), the two measures of extraction described in Section 2 and Appendix D.1.

ID	Model	R-1	R-2	R-L	BS	MC_{EXT}	MC_{INPUT}	# Sents	# Words
1	**QT_{asp}**	31.61	10.24	22.64	87.93	28.31	**47.87**	5.79	46.80
2	$AceSum_{ext}$	35.09	12.10	27.15	88.95	**29.10**	44.82	1.99	29.56
3	$HERCULES_{ext}$	28.57	7.91	19.93	87.76	10.30	23.92	4.01	56.40
4	$HERCULES_{abs}$	33.56	10.04	25.34	89.59	23.87	39.90	3.99	33.85
5	AceSum	**36.38**	**12.65**	**29.08**	**89.65**	26.59	40.79	2.21	27.84
6	MISTRAL	28.67	**7.54**	20.06	88.19	25.36	59.59	5.00	74.70
7	*w/ Self-Refine*	23.98	5.59	16.77	87.32	23.49	63.66	5.34	124.18
8	*w/ MC-Refine*	**30.50***	7.52	**22.00***	**88.51***	**34.63***	**78.79***	2.73	55.16
9	LLAMA3.1	**30.26**	7.05	21.96	**88.64**	31.61	74.03	1.78	47.70
10	*w/ Self-Refine*	29.07	5.92	20.82	88.10	17.89	52.37	1.64	52.66
11	*w/ DCR*	25.27	5.57	20.31	88.10	43.35	60.43	1.15	21.48
12	*w/ MC-Refine*	30.23	**7.23**	**22.48**	88.55	**55.51***	**76.54***	1.94	54.16
13	LLAMA3.1-70B	31.42	7.70	22.93	88.95	29.39	66.70	1.69	48.12
14	*w/ Self-Refine*	29.54	6.46	21.68	88.57	19.63	54.19	1.26	46.19
15	*w/ DCR*	26.70	5.98	21.49	88.78	46.30	66.45	1.20	22.78
16	*w/ MC-Refine*	**32.88***	**8.41***	**24.85***	**89.30***	**50.06***	**78.38***	1.24	39.71

Table 18: Experimental results on SPACE. All LLM results are averaged over three runs, and * means the score is significantly better than the baseline models within each block. Gray rows indicate the baseline models, The best scores for each backbone and the prior baselines are **bold**.

Model	Summary
Human References	['This charming hotel is located within steps of all the major sights', 'The hotel is perfectly situation a few steps away from all major sights to see in Rome.', 'Its in a great location with an available walking tour of Rome.']
AceSum	the location is ideally located just 2 minutes walk from piazza navona and a 15 minute walk to st. peter's and the roman forum.
LLama3.1-70B	Hotel Navona's location is excellent, within walking distance to major attractions like Piazza Navona, Pantheon, Vatican, and Spanish Steps, making it an ideal base for exploring Rome, with many restaurants and cafes nearby, and public transportation easily accessible.
LLama3.1-70B w/ Self-Refine	Hotel Navona is nestled in the heart of Rome's historic center, steps away from Piazza Navona and the Pantheon, with easy access to Vatican, Spanish Steps, and nearby restaurants, shops, and public transportation, making it an ideal base for exploring Rome.
LLama3.1-70B w/ DCR	Hotel Navona's location is ideal for exploring Rome, with many restaurants and cafes nearby, and public transportation easily accessible.
llama3.1-70B w/ MC-Refine	Hotel Navona's location is excellent, being only 2 minutes from the Pantheon and Piazza Navona and a 15 minute walk to St. Peter's and the Roman Forum, making it an ideal base for exploring Rome.

Table 19: Example Sumamries of different approaches on SPACE.