# CARPAS: Towards Content-Aware Refinement of Provided Aspects for Summarization in Large Language Models

**Yong-En Tian and Yu-Chien Tang and An-Zi Yen and Wen-Chih Peng**
National Yang Ming Chiao Tung University
bryanttian.cs12@nycu.edu.tw, tommytyc.cs10@nycu.edu.tw

## Abstract

Aspect-based summarization has attracted significant attention for its ability to generate more fine-grained and user-aligned summaries. While most existing approaches assume a set of predefined aspects as input, real-world scenarios often present challenges where these given aspects may be incomplete, irrelevant, or entirely missing from the document. Users frequently expect systems to adaptively refine or filter the provided aspects based on the actual content. In this paper, we initiate this novel task setting, termed Content-Aware Refinement of Provided Aspects for Summarization (CARPAS), with the aim of dynamically adjusting the provided aspects based on the document context before summarizing. We construct three new datasets to facilitate our pilot experiments, and by using LLMs with four representative prompting strategies in this task, we find that LLMs tend to predict an overly comprehensive set of aspects, which often results in excessively long and misaligned summaries. Building on this observation, we propose a preliminary subtask to predict the number of relevant aspects, and demonstrate that the predicted number can serve as effective guidance for the LLMs, reducing the inference difficulty, and enabling them to focus on the most pertinent aspects. Our extensive experiments show that the proposed approach significantly improves performance across all datasets. Moreover, our deeper analyses uncover LLMs' compliance when the requested number of aspects differs from their own estimations, establishing a crucial insight for the deployment of LLMs in similar real-world applications.[1]

## 1 Introduction

Aspect-based summarization (ABS) offers tailored summaries that focus on specific topics or facets within documents, enabling users to quickly extract targeted information. This fine-grained summarization approach has proven highly valuable across multiple domains, such as customer reviews (Kunneman et al., 2018), scientific papers (Takeshita et al., 2024), and financial earnings call transcripts (Tian et al., 2025). While existing ABS methods have shown strong performance, they typically rely on a predefined and accurate set of aspects as input. These approaches can broadly be categorized into two main types. The first type adopts a two-stage methodology, where the initial stage is dedicated to generating aspect-based elements, such as identifying key sentences (Hayashi et al., 2021; Amar et al., 2023) or relevant keywords (Zhu et al., 2021; Tan et al., 2020) for each aspect. In the second stage, this identified and aspect-specific content is then fed into a summarization model to produce the final summary. Alternatively, the second type employs an end-to-end approach, where the model directly generates the summary using the entire source document and aspects as input (Guo and Vosoughi, 2024b; Takeshita et al., 2024). However, this inherent reliance on fixed or pre-extracted aspects presents a significant limitation when the given aspects are incomplete, irrelevant, or entirely absent from the document content, a common challenge

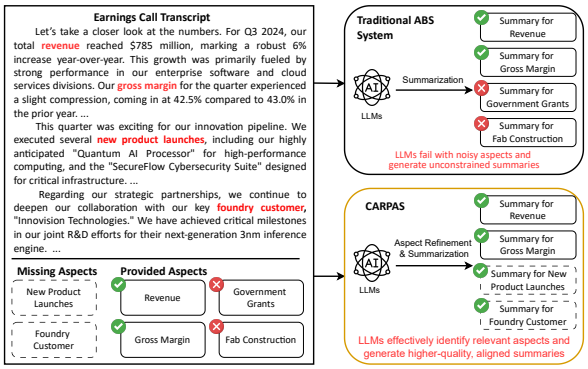



Figure 1: CARPAS aims to refine the provided aspects to generate more aligned summaries, a task that is challenging for previous ABS approaches.

[1] Our code and datasets will be publicly available to support future research on CARPAS.

in real-world applications. For instance, in the scenario illustrated in Figure 1, the document's ground-truth aspects include `Revenue`, `Gross Margin`, `New Product Launches`, and `Foundry Customer`. Yet, the initially provided aspects might be `Revenue`, `Gross Margin`, `Government Grants`, and `Fab Construction`. Traditional ABS approaches would attempt to generate summaries for all the provided aspects, regardless of their actual relevance. Ideally, a system should not only filter out the irrelevant provided aspects (e.g., `Government Grants` and `Fab Construction`) but also identify the missing ones that are present in the document (such as `New Product Launches` and `Foundry Customer`). This gap motivates us to develop a new task: Content-Aware Refinement of Provided Aspects for Summarization (CARPAS).

Given the remarkable advancements of LLMs in recent years (Minaee et al., 2024; Dubey et al., 2024; Hurst et al., 2024; Bi et al., 2024), we conduct preliminary experiments to address this challenge using an end-to-end approach. In this setup, LLMs are tasked with first predicting the correct aspects, and then generating corresponding summaries based on these predicted aspects. We explore four representative prompting strategies: direct prompting (Brown et al., 2020), Chain-of-Thought prompting (Wei et al., 2022), Chain-of-Thought with self-consistency (Wang et al., 2023), and Self-Refine (Madaan et al., 2023). However, our initial results indicate that relying solely on the LLMs' inherent capabilities with these prompting strategies achieves an average BERTScore of only 57.36% on the earnings call transcripts dataset, an unsatisfactory performance for accurate content-aware aspect refinement in real-world deployment. In this pilot study, we observe that LLMs tend to predict a more comprehensive set of aspects, often leading to the generation of overly extensive summaries. This behavior likely stems from LLMs attempting to cover the entire document by generating an excessive number of aspects.

To mitigate this, we introduce a prior subtask: predicting the number of relevant aspects. This predicted number is then provided to the LLM as additional information, offering a stronger foundation for its aspect prediction and thereby reducing the difficulty of the inference process. As there is no existing dataset specifically tailored for this task and the predominant cost to annotate one, we leverage LLM to generate synthetic data for two distinct domains: earnings call transcripts and COVID-19 press conference materials. Additionally, to assess its real-world applicability, we augment real-world earnings call transcript data to further evaluate the practicality and effectiveness of our subtask. Our experiments in these three datasets and four prompting strategies demonstrate that our proposed approach significantly improves performance in all experimental settings.

In sum, the main contributions of this paper are threefold: (1) We introduce a practical task CARPAS to address the proliferating ABS task in recent years, and release the corresponding code and datasets to facilitate future research. (2) We empirically validate that LLMs with popular prompting strategies cannot consistently generate great summaries in CARPAS, and thus design a simple yet effective method that can significantly enhance the performance by predicting the number of relevant aspects in advance. (3) Quantitative results highlight the validity of integrating the aspect-counts prediction into CARPAS. Our findings further reveal LLMs' unconditional faith in the counts of initially provided aspect sets, offering a reflective view to direct employment of LLMs in similar industrial settings.

## 2 Related Work

### 2.1 Aspect-Based Summarization Datasets

Early ABS datasets assume predefined aspects, often derived from structural cues like Wikipedia section titles (Hayashi et al., 2021), manually curated labels in scientific domains (e.g., purpose, methodology) (Takeshita et al., 2024; Meng et al., 2021), or fixed categories for news and meetings (Ahuja et al., 2022; Deng et al., 2023). More recent work supports open or dynamic aspects to better reflect real-world scenarios where aspects may be implicit or user-defined. For instance, OpenAsp (Amar et al., 2023) and OASum (Yang et al., 2023b) create open-domain datasets via manual annotation, while other benchmarks use entities as proxy aspects (Maddela et al., 2022) or discover latent aspects in unordered text (Guo and Vosoughi, 2024a). Despite these advances, existing datasets still assume aspects are either explicitly provided or reliably extracted from the document, an assumption that fails in flexible, user-guided scenarios. Crucially, no dataset is designed for the CARPAS task, where models must dynamically refine, filter, or reinterpret potentially noisy, incomplete, or irrelevant aspect sets based on the document content.

This gap motivates our construction of synthetic datasets to pioneer this field.

### 2.2 Aspect-Based Summarization Methods

The goal of ABS is to produce summaries targeted to specific aspects or facets of the content. These methods can broadly be categorized into two main types: two-stage and end-to-end. Two-stage approaches first identify aspect-relevant content and then summarize it. Early work grouped sentences by aspect (Angelidis and Lapata, 2018; Hayashi et al., 2021), while others identify key sentences or keywords (Tan et al., 2020; Zhu et al., 2021) or cluster aspect-indicative tokens in an unsupervised manner (Li et al., 2023). The extracted content is then passed to a summarization model. In contrast, end-to-end methods use a unified framework, conditioning generation on aspect-related signals (e.g., tokens, prompts) to jointly learn relevance and summary formulation (He et al., 2022; Ahuja et al., 2022). For example, some models employ multi-objective training for dynamic aspects (Guo and Vosoughi, 2024b) or latent structures to bypass the need for labeled data (Frermann and Klementiev, 2019). Recent research increasingly applies LLMs to ABS tasks, leveraging their powerful generation capabilities. This includes fine-tuning open-source LLMs like LLaMA2 and Mistral (Mullick et al., 2024), prompting models like ChatGPT (Yang et al., 2023a), and hybrid approaches that combine retrieval with LLM generation to improve factual accuracy (Feng et al., 2025). LLM-driven workflows have also been used in specific domains like summarizing financial transcripts (Tian et al., 2025). However, all these existing methods, including LLM-based pipelines, rely on a pre-defined and accurate set of input aspects. Our proposed CARPAS task introduces a new setting that challenges models to dynamically refine imperfect aspect candidates, a scenario unaddressed by prior ABS research.

## 3 Method

### 3.1 Dataset Construction

**Synthetic data.** Previous work has explored using LLM-synthesized data for novel tasks to address the high cost of real-world data collection (Frermann and Klementiev, 2019; Whitehouse et al., 2023; Nadas et al., 2025). These results demonstrate that synthetic data can yield findings similar to those from human-annotated datasets. Inspired by this, we construct a synthetic dataset by prompting Gemini-2.0-Flash (Comanici et al., 2025) to generate data for two distinct categories: earnings call transcript (denoted as ECT), and the daily Central Epidemic Command Center COVID-19 press conference (denoted as COVID-19-PC). These domains are representative of complex real-world scenarios characterized by multiple speakers discussing numerous topics in an interleaved manner.

| Dataset | #Aspects Per Doc. | | | | | Avg. #Tokens | |
|---|---|---|---|---|---|---|---|
| | 4 | 5 | 6 | 7 | 8 | Aspect Summary | Doc. |
| ECT | 29 | 28 | 29 | 32 | 27 | 97.64 | 2979.72 |
| COVID-19-PC | 31 | 34 | 33 | 31 | 31 | 86.43 | 2299.40 |
| RW-ECT | 19 | 17 | 20 | 20 | 17 | 105.54 | 3745.00 |

Table 1: Statistics of three datasets grouped by number of aspects and the average number of tokens in each document.

For ECT, we utilize a default aspect set from (Tian et al., 2025), which comprises 18 key aspects relevant to earnings call transcripts. For COVID-19-PC, we refer to the transcript structure of government COVID-19 press conference[2] and define 26 aspects. We first construct transcript templates based on typical structures observed in real transcripts for both domains, which include sections such as opening remarks, discussion, topics highlights, and a Q&A session, ensuring a realistic document flow. Second, we randomly sample a total number of aspects (ranging from 4 to 8) from the default aspect set. The LLM is then tasked with writing a content paragraph with specific and plausible details for each sampled aspect. These aspect-specific contents are subsequently combined with the transcript template by prompting the LLM to synthesize them into a transcript. Finally, the LLM is instructed to create an aspect-based summary for each paragraph, which serves as the ground-truth data for our experiments. This controlled generation process ensures that each synthesized document is constructed around a precise number of ground-truth aspects, creating a well-defined benchmark for our experiments. As a result, a total of 145 documents are generated for ECT and 160 for COVID-19-PC.

**Real-world data.** We collect earnings call transcripts from five leading semiconductor manufacturing companies, spanning from 2019 to 2024,

[2] https://youtube.com/playlist?list=PLSHckwvN6OpIUHJPHeXB1rxqyWhJozfxR

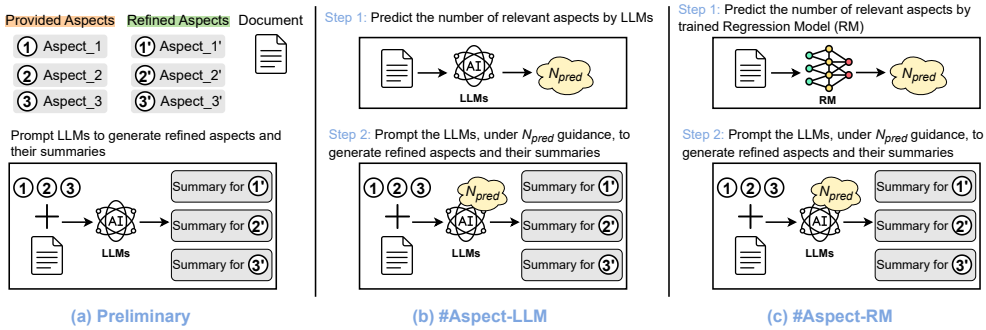



Figure 2: Overall workflows of three approaches for CARPAS. The Preliminary method (a) directly refines the provided aspects using prompting strategies. The #Aspect-LLM method (b) adds aspect number prediction via LLMs. The #Aspect-RM method (c) incorporates aspect number prediction from a trained regression model.

as reference sources for the detailed content paragraphs for each sampled aspect. This augmentation process results in 93 real-world ECT documents (denoted as RW-ECT).

Table 1 summarizes the key statistics of all datasets, including document counts, aspect numbers, document tokens, and aspect summary tokens. The prompts for data construction and examples of aspect sets for each category are detailed in Appendix C.

### 3.2 Preliminary Experiments

In the Preliminary Experiments, as stated in Figure 2 (a), we employ four representative prompting strategies to tackle the CARPAS task: direct prompting (Brown et al., 2020), Chain-of-Thought prompting (Wei et al., 2022), Chain-of-Thought with self-consistency (Wang et al., 2023), and Self-Refine (Madaan et al., 2023). These strategies are utilized to refine the provided aspects and subsequently generate the corresponding aspect-based summaries. For each strategy, the LLMs are given the full document and the provided aspects (which could include irrelevant or missing ones). The goal is to produce refined versions of the provided aspects that are relevant to the document, followed by a summary for each refined aspect. The four strategies are defined as follows:

**Direct Prompting (DP)**: This approach involves providing the LLM with a straightforward instruction that describes the task, requesting the refined aspects and summaries directly as the final output.

**Chain-of-Thought Prompting (CoT)**: This strategy prompts the LLM to first generate a step-by-step explanation of its reasoning for refining the provided aspects (i.e., why specific aspects are retained, modified, removed, or added) before generating the final answer.

**Chain-of-Thought with Self-Consistency (CoT-SC)**: As an extension of CoT, this approach generates multiple, diverse reasoning paths for the same task. The final, most consistent answer is then determined by a concluding LLM prompt that instructs the model to evaluate all generated paths and select the most coherent and logical one, which enhances the robustness of the results.

**Self-Refine (SR)**: This approach adopts an iterative, agentic framework with a conditional workflow. The first agent evaluates the provided aspects to determine whether refinement is necessary. If refinement is required, the second agent performs the necessary modifications and then returns the revised aspects to the first agent for re-evaluation. This process may repeat until the first agent determines that no further refinement is needed. Once the aspects are finalized, the third agent generates summaries based on them.

### 3.3 Aspect Number Prediction

During the Preliminary Experiments, we observe that LLMs often struggle to generate summaries that accurately reflect the document content. As shown in Table 2 and Figure 3, the final summaries achieve unsatisfactory BERTScore and ROUGE-L (approximately 60% and 30%), primarily due to a mismatch between the total number of predicted aspects and the ground-truth annotations. To address this limitation and guide the LLMs more effectively, we introduce a dedicated subtask: Aspect Number Prediction. We hypothesize that the suboptimal performance observed in preliminary experiments stems from the LLMs' tendency to comply with the total number of provided aspects, and with proper guidance, such as prompting the model with an accurate estimate of the number of aspects, LLMs are capable of correctly predicting

the aspect set. For this subtask, the input is the full document, and the output is a single integer representing the estimated number of relevant aspects within that document. We explore two approaches for this prediction, and the overall workflows are shown in Figure 2 (b), (c).

#### 3.3.1 #Aspect-LLM

The first approach involves directly prompting LLMs to predict the number of relevant aspects based on the document content by asking them "Estimate how many distinct, non-overlapping aspects or key themes are present". This method aims to test whether LLMs themselves can precisely predict the aspect number.

#### 3.3.2 #Aspect-RM

The second approach focuses on training a regression model composed of an embedding model followed by a linear layer to predict the correct aspect number. Specifically, we use `Qwen3-Embedding-0.6B` (Zhang et al., 2025) as the embedding model and train the regression model by minimizing the Mean Absolute Error (MAE) loss. We perform this fine-tuning on the training split of our synthetic datasets, where the ground-truth aspect number for each document is known from its generative process.

#### 3.3.3 Combined with Aspect Refinement

The predicted aspect number $N_{pred}$ from either of these methods is then applied to the four prompting strategies, serving as a guidance for the LLMs in the subsequent aspect refinement and summarization steps. LLMs are expected to output refined aspects by tasking them with following prompt: "This document likely contains around $N_{pred}$ aspects. Please identify the $N_{pred}$ most relevant aspects and provide a concise summary for each." This allows for precise control over the output granularity, and can further prevent the model from generating overly detailed or sparse aspect sets, an issue that often arises when the LLMs misinterpret the correct aspects.

## 4 Experiments

### 4.1 Experimental Setup

#### 4.1.1 LLMs and Prompting Strategies

In our experiments, we utilize Gemma-3 12B (G3-12B), Gemma-3 27B (G3-27B) (Kamath et al., 2025), GPT-4o-mini (4o-mini), and GPT-4o (4o) (Hurst et al., 2024) as the LLMs.[3] For all LLM interactions, the temperature is set to 0.7 to balance creativity and consistency in generation. The maximum number of iterations for Self-Refine prompting is set to 16.

#### 4.1.2 Datasets

The train-test splitting of each dataset is 75:70 for ECT, 85:75 for COVID-19-PC, and 60:33 for RW-ECT. As for the provided aspects, we simulate real-world scenarios with three different settings, where $y$, $n$ denote the correct and incorrect number of aspects: (1) Completely incorrect aspects: $(y, n) \in \{(0, 2), (0, 4), (0, 6)\}$. (2) Completely correct aspects: $(y, n) \in \{(2, 0), (4, 0), (6, 0)\}$. (3) A mix of correct and incorrect aspects: $(y, n) \in \{(2, 2), (4, 4), (6, 6)\}$. For simplicity, we adopt a shorthand notation in the form of `y2n2` to represent $(y, n) = (2, 2)$ in the following sections.

#### 4.1.3 Implementation Details of #Aspect-RM

The Regression Model is trained using the training splits of each dataset. Through our experiments, we apply LoRA (Hu et al., 2022) for fine-tuning, which effectively reduces memory consumption and enables more efficient model training, and we use AdamW (Loshchilov and Hutter, 2019) as the optimizer. The model is trained for 30 epochs with a batch size of 1, a learning rate of 2e-5, and a fixed random seed of 42.

#### 4.1.4 Evaluation Metrics

To measure the quality of the generated summary, we use BERTScore (Zhang et al., 2020) along with ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) as the main evaluation metrics, following (Guo and Vosoughi, 2024a). The BERT-based model for BERTScore is `deberta-v3-large` (He et al., 2023). In addition, we assess the accuracy of aspect identification by calculating the absolute difference in the number of aspects between the reference and generated summaries (denoted as #AbsAspDiff, the lower the better). More detailed evaluation metrics and results are described in the Appendix A.

### 4.2 Quantitative Results

Table 2 presents the overall performance of all LLMs and prompting strategies across synthetic (ECT and COVID-19-PC) and real-world (RW-ECT) datasets. It can be seen that #Aspect-RM sig-

[3] In addition, we also experiment with several reasoning models and the results are detailed in the Appendix B.

| Prompting Strategy | ECT | | | | | | | | COVID-19-PC | | | | | | | | RW-ECT | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G3-12B | | G3-27B | | 4o-mini | | 4o | | G3-12B | | G3-27B | | 4o-mini | | 4o | | G3-12B | | G3-27B | | 4o-mini | | 4o | |
| | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L | BS | R-L |
| **Preliminary** | | | | | | | | | | | | | | | | | | | | | | | | |
| DP | 56.1 | 32.6 | 62.6 | 33.6 | 50.9 | 23.7 | 62.7 | 31.7 | 56.9 | 33.8 | 64.5 | 35.4 | 47.9 | 20.9 | 57.6 | 27.8 | 54.9 | 29.1 | 62.0 | 30.1 | 48.3 | 19.8 | 61.0 | 27.9 |
| CoT | 58.1 | 32.8 | 63.5 | 34.3 | 45.3 | 21.0 | 61.6 | 28.3 | 55.2 | 31.8 | 64.2 | 35.3 | 40.9 | 17.6 | 60.0 | 27.0 | 54.6 | 29.5 | 62.3 | 31.6 | 43.0 | 18.3 | 59.8 | 25.6 |
| CoT-SC | 63.6 | 34.1 | 63.1 | 31.1 | 48.5 | 23.4 | 63.7 | 30.6 | 66.2 | 38.1 | 65.2 | 34.7 | 46.2 | 21.4 | 62.2 | 30.4 | 61.2 | 30.6 | 62.1 | 29.0 | 46.8 | 20.2 | 62.5 | 28.2 |
| SR | 56.6 | 32.2 | 51.7 | 28.3 | 53.2 | 25.7 | 56.6 | 26.6 | 60.2 | 35.3 | 52.2 | 29.6 | 52.7 | 26.7 | 56.0 | 28.9 | 55.4 | 29.2 | 51.4 | 26.6 | 51.4 | 24.0 | 54.7 | 25.9 |
| **#Aspect-LLM** | | | | | | | | | | | | | | | | | | | | | | | | |
| DP | 53.5 | 30.7 | 61.6 | 32.0 | 50.9 | 23.0 | 56.3 | 27.7 | 54.2 | 32.2 | 61.0 | 33.2 | 55.1 | 23.6 | 56.2 | 26.5 | 56.0 | 29.1 | 59.0 | 27.9 | 51.1 | 20.7 | 58.7 | 25.7 |
| CoT | 59.5 | 32.8 | 62.8 | 32.6 | 52.3 | 23.5 | 55.8 | 25.1 | 58.8 | 34.0 | 61.4 | 33.6 | 53.4 | 22.5 | 55.8 | 24.5 | 60.5 | 31.1 | 60.4 | 29.7 | 52.9 | 21.6 | 58.2 | 24.1 |
| CoT-SC | 59.2 | 31.5 | 60.7 | 29.6 | 56.1 | 26.0 | 60.8 | 28.6 | 55.6 | 31.7 | 59.9 | 31.2 | 55.3 | 24.5 | 56.1 | 26.8 | 56.3 | 26.6 | 57.7 | 25.7 | 51.6 | 21.6 | 58.3 | 25.3 |
| SR | 56.1 | 32.0 | 53.2 | 29.0 | 51.7 | 26.6 | 55.3 | 28.7 | 54.0 | 30.8 | 60.4 | 34.3 | 54.0 | 26.3 | 57.0 | 28.1 | 45.4 | 21.9 | 58.4 | 29.3 | 51.0 | 22.5 | 59.1 | 26.2 |
| **#Aspect-RM** | | | | | | | | | | | | | | | | | | | | | | | | |
| DP | 70.7 | **38.6** | 72.9 | 36.7 | 59.7 | 26.7 | 71.5 | 34.0 | 69.7 | 39.8 | **74.4** | 38.9 | 64.2 | 26.5 | 71.0 | 32.4 | 68.9 | **34.7** | 73.0 | 33.9 | 59.4 | 23.6 | 71.4 | 31.2 |
| CoT | 71.2 | 38.3 | **73.5** | **37.7** | 61.4 | 27.3 | 69.3 | 30.3 | 70.1 | 39.4 | 74.2 | 39.4 | 61.6 | 25.4 | 69.0 | 30.1 | 68.0 | 34.0 | **73.1** | **34.9** | 59.5 | 23.8 | 68.1 | 27.7 |
| CoT-SC | **72.0** | 35.7 | 71.9 | 33.6 | 69.9 | 31.2 | 71.0 | 32.3 | **73.9** | **41.1** | 74.0 | 37.2 | 68.7 | 29.5 | 72.4 | 33.7 | **71.5** | 32.6 | 71.8 | 31.4 | 69.2 | 27.8 | 71.2 | 30.1 |
| SR | 65.2 | 34.2 | 70.6 | 37.0 | **70.6** | **33.5** | **71.9** | **34.4** | 66.9 | 38.1 | 72.4 | **40.3** | **71.2** | **34.1** | **73.1** | **35.9** | 63.0 | 30.4 | 70.3 | 34.7 | **70.2** | **30.7** | **71.7** | **31.7** |

Table 2: Results (×100%) on synthetic (ECT and COVID-19-PC) and real-world (RW-ECT) datasets, with BS standing for BERTScore, and R-L standing for ROUGE-L. The highest results for each metric are in **boldface**, while the second-best are underlined.

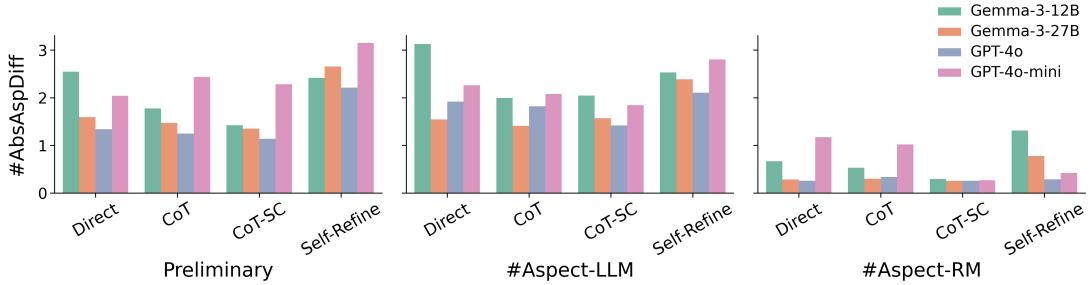



Figure 3: Aspect count error (#AbsAspDiff) across different methods and prompting strategies on ECT.

nificantly improves both BERTScore and ROUGE-L within Preliminary Experiments settings, and consistently outperforms #Aspect-LLM. As shown in Figure 4, this superior performance is consistent across all provided aspect settings on ECT. Compared to the baseline in Preliminary Experiments, #Aspect-RM achieves an average improvement of 25.4% in BERTScore and 19.8% in ROUGE-L on the ECT dataset. Even larger gains are observed on the COVID-19-PC dataset, with BERTScore and ROUGE-L increasing by 30.1% and 24.4%, respectively. This highlights the impact of providing accurate numerical guidance as a prerequisite for generating high-quality, aspect-aligned summaries. On RW-ECT, #Aspect-RM continues to deliver substantial gains, with BERTScore and ROUGE-L increasing by 23.4% and 15.9%, demonstrating that our proposed subtask is not only effective on synthetic data but also robustly enhances summarization quality on complex, real-world financial documents.

The primary driver of this performance gain is the model's enhanced ability to identify the correct number of aspects. As illustrated in Figure 3, #Aspect-RM achieves the lowest #AbsAspDiff by a large margin, reducing by an average of 71.7% in ECT and 82.4% in COVID-19-PC compared to the Preliminary Experiments. In contrast, simply prompting the LLMs for an aspect number, as done in #Aspect-LLM, fails to yield improvements over the Preliminary Experiments baseline due to inaccurate aspect number estimation. Figure 3 also highlights the severity of this issue in baseline methods: in both the Preliminary Experiments and #Aspect-LLM results, the generated aspect count deviates from the ground truth by at least one aspect on average, and sometimes by as many as two or three. This consistent overestimation leads to the generation of superfluous aspects and summaries, underscoring the benefit of explicit aspect number guidance to LLMs in avoiding both the omission of relevant aspects and the hallucination of irrelevant ones.

Another interesting observation is that once guided by #Aspect-RM, more sophisticated prompting strategies will perform better. Specifically, GPT-based models achieve their best performance with Self-Refine, an agentic strategy that

| | Models | Avg. Step | BS(↑) | #AbsAspDiff(↓) |
|---|---|---|---|---|
| Preliminary | G3-12B | 3.215 | 0.602 | 1.899 |
| | G3-27B | 2.516 | 0.522 | 2.492 |
| | 4o-mini | 3.582 | 0.527 | 2.204 |
| | 4o | 3.363 | 0.560 | 2.034 |
| #Aspect-LLM | G3-12B | 5.864 | 0.540 | 2.932 |
| | G3-27B | 3.653 | 0.603 | 1.803 |
| | 4o-mini | 5.443 | 0.540 | 2.480 |
| | 4o | 3.997 | 0.570 | 1.930 |
| #Aspect-RM | G3-12B | 6.145 | 0.669 | 1.324 |
| | G3-27B | 3.556 | 0.724 | 0.563 |
| | 4o-mini | 5.084 | 0.712 | 0.250 |
| | 4o | 3.852 | 0.731 | 0.099 |

Table 3: Results of Self-Refine on COVID-19-PC, showing average thinking steps, BERTScore, and #AbsAspDiff.

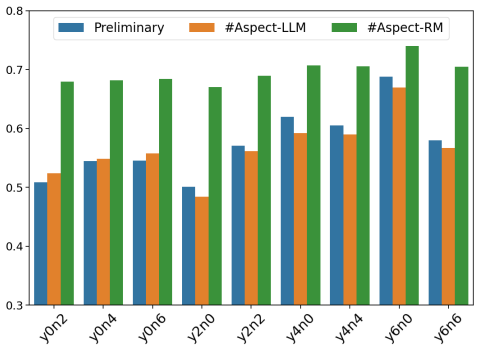



Figure 4: Average BERTScore results on ECT across different provided aspect settings.

leverages autonomous decision-making. On average, Self-Refine and CoT-SC emerge as the top-performing strategies under our framework, in line with the fact that self-reflection and multi-agent system excel in complex generation tasks (Asai et al., 2024; Shinn et al., 2023).

### 4.3 Discussion

**Effect of Self-Refine Thinking Steps.** To further probe the operational dynamics of Self-Refine, we analyze its average thinking steps (Avg. Step) on the COVID-19-PC dataset and summarize the results in Table 3. Our findings show that both #Aspect-LLM and #Aspect-RM guidance encourage LLMs to take more thinking steps; however, this does not consistently lead to better performance. This indicates that accurately estimating the number of aspects is more crucial than simply increasing the number of refinement steps. Surprisingly, our analysis also uncovers an inverse relationship between model size and reasoning effort: smaller models consistently exhibit a higher average step than their larger counterparts. Similar trends are observed on the other datasets as well, providing insights that larger model might sometimes overthink, making unnecessary edits to the aspects and summaries. The detailed results on other datasets and LLMs are provided in Table 11.

**Quality of Predicted Aspect Set.** To delve into the quality of the predicted aspect set, we further evaluate the factuality and relevance between the predicted aspects and the document content by prompting `Qwen3-Next-80B-A3B-Instruct` on a 1–5 Likert scale. In addition, we assess the similarity between the predicted and ground-truth aspects using a weighted BERTScore, $\text{BS}_\text{w}$:

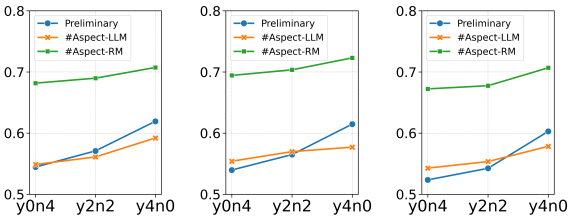
$$BS_w = \frac{1}{n + |n - m|} \sum_{i=1}^{n} BS(a^i_{\text{pred}}, a^i_{\text{ground}}) \quad (1)$$

where $m$ and $n$ denote the number of predicted and ground-truth aspects, respectively, and $a^i_{\text{pred}}$ and $a^i_{\text{ground}}$ are the $i$-th aspects in a matched pair. This metric penalizes the over- or under-estimation of aspect counts and can thus provide a more accurate assessment of the aspect set quality for each method. For the paired computation of BERTScore, we apply the Hungarian Algorithm (Kuhn, 1955) to find the optimal pairing of aspects that maximizes the total similarity score. The results, averaged over four prompting strategies, are presented in Table 4. Overall, we observe that #Aspect-RM achieves superior factuality, relevance, and $\text{BS}_\text{w}$ across all datasets and models, indicating that providing an accurate aspect count estimation to LLMs can effectively improve the quality of the predicted aspects, hence benefiting the downstream ABS task.

**Impact of Provided Aspects.** We study the detailed breakdown of how the quality of initial provided aspects influences summarization performance, and conduct experiments with the three settings: completely incorrect (y0n4), a mix of correct and incorrect (y2n2), and completely correct (y4n0). The results are reported in Figure 5. A clear trend emerges across all approaches: summarization quality, as measured by average BERTScore, improves as the proportion of correct aspects in the initial input increases. This highlights that noisy input aspects might potentially lead to inconsistencies and ultimately hinder the summarization performance of LLMs. In addition, Figure 5 reveals the robustness of the Regression Model approach. While the performance of the Preliminary Experiments is highly sensitive to the quality of the initial provided aspects, our method maintains a

| Dataset | Method | Gemini-2.5-Flash | | | Gemma-3-12B | | | Gemma-3-27B | | | GPT-4o-mini | | | GPT-4o | | | O3-mini | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Fact. | Rel. | $BS_w$ | Fact. | Rel. | $BS_w$ | Fact. | Rel. | $BS_w$ | Fact. | Rel. | $BS_w$ | Fact. | Rel. | $BS_w$ | Fact. | Rel. | $BS_w$ |
| ECT | Preliminary | **4.67** | 4.10 | 0.46 | 4.66 | 3.82 | 0.47 | **4.69** | 3.82 | 0.43 | 4.62 | 3.32 | 0.43 | 4.71 | 4.04 | 0.47 | **4.79** | 3.80 | 0.45 |
| | #Aspect-LLM | **4.67** | 4.00 | 0.38 | 4.66 | **3.89** | 0.43 | **4.69** | 3.78 | 0.43 | 4.62 | 3.64 | 0.46 | **4.72** | 3.85 | 0.46 | 4.78 | 4.13 | 0.48 |
| | #Aspect-RM | 4.65 | **4.22** | **0.54** | **4.67** | 3.87 | **0.50** | 4.68 | **4.01** | **0.49** | **4.67** | **3.83** | **0.52** | 4.71 | **4.12** | **0.55** | **4.79** | **4.19** | **0.54** |
| COVID-19-PC | Preliminary | 4.97 | 4.10 | 0.51 | **4.99** | 4.16 | 0.54 | 4.98 | 4.19 | 0.50 | 4.88 | 3.49 | 0.51 | **5.00** | 4.19 | 0.55 | 4.99 | 4.04 | 0.52 |
| | #Aspect-LLM | **5.00** | 4.07 | 0.42 | **4.99** | 4.17 | 0.48 | **4.99** | 4.29 | 0.47 | 4.97 | 4.06 | 0.53 | 4.99 | 4.16 | 0.50 | 4.99 | 4.37 | 0.55 |
| | #Aspect-RM | **5.00** | **4.47** | **0.61** | **4.99** | **4.38** | **0.61** | **4.99** | **4.48** | **0.57** | **4.98** | **4.23** | **0.65** | 4.99 | **4.44** | **0.64** | **5.00** | **4.52** | **0.62** |
| RW-ECT | Preliminary | 3.75 | 3.89 | 0.45 | **3.79** | **3.58** | 0.47 | **3.89** | 3.54 | 0.42 | **3.98** | 3.08 | 0.45 | 3.89 | 3.80 | 0.48 | **4.08** | 3.56 | 0.44 |
| | #Aspect-LLM | 3.65 | 3.88 | 0.36 | 3.76 | 3.57 | 0.39 | 3.83 | 3.68 | 0.40 | 3.83 | 3.54 | 0.42 | 3.89 | 3.71 | 0.45 | 4.01 | 3.83 | 0.45 |
| | #Aspect-RM | **3.76** | **4.01** | **0.54** | **3.79** | **3.58** | **0.49** | **3.89** | **3.72** | **0.49** | 3.93 | **3.67** | **0.52** | **3.90** | **3.92** | **0.54** | 3.98 | **3.96** | **0.54** |

Table 4: The factuality (Fact.), relevance (Rel.), and $BS_w$ of predicted aspects across three datasets and models.




(a) ECT (b) COVID-19-PC (c) RW-ECT

Figure 5: Average BERTScore across three provided aspect settings on the three datasets.

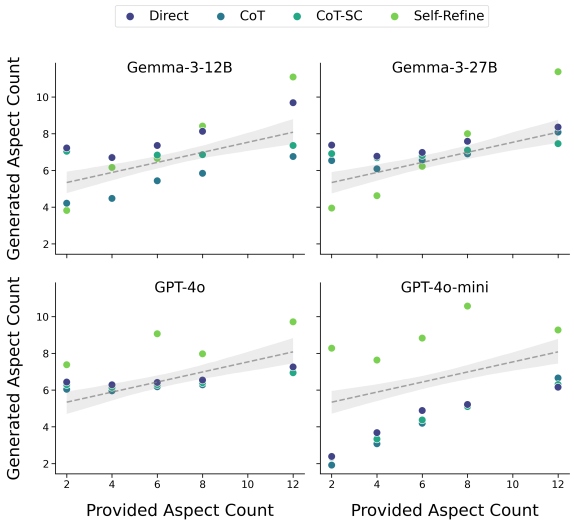



Figure 6: Relationship between the provided aspect count and the generated aspect count on ECT under the Preliminary Experiments.

high BERTScore even in the most challenging scenario (y0n4). This demonstrates the effectiveness of #Aspect-RM in filtering out noise and focusing on content-relevant aspects, a critical capability for real-world applications where provided aspects are often imperfect.

Furthermore, a closer examination of the total number of generated aspects, as illustrated in Figure 6, unveils a positive correlation with the total number of provided aspects. In the Preliminary Experiments, we observe that LLMs may exhibit a compliance phenomenon, where they tend to directly generate a similar number of summaries as the total number of provided aspects without sufficient critical thinking or effectively refining the aspect set. This behavior can negatively impact summarization quality, as the model is compelled to produce content for irrelevant or non-existent aspects. Such overgeneration leads to hallucinations and factual inaccuracies, ultimately degrading the overall BERTScore, as evidenced in Figure 5.

## 5 Conclusion

This paper introduces CARPAS, a novel and practical task setting that addresses the challenge of handling imperfect input aspects in real-world ABS problems. To tackle this, we propose a simple yet highly effective two-stage framework that first predicts the number of relevant aspects by a regression model, and then uses this information to guide LLMs for the aspect refinement and following summarization in a content-aware manner. Experiments on three different datasets show that our approach consistently outperforms popular prompting strategies, validating our intuition that providing explicit numerical guidance to LLMs can improve their performance in CARPAS. We believe our results provide useful insights for similar aspect-finding applications, and multiple future research directions can be further explored, including applying our approach to areas such as structured data-to-text generation or few-shot dialogue systems, further extending the impact of our findings.

## Limitations

While our proposed framework demonstrates strong results on the newly introduced CARPAS task, several limitations remain, pointing to directions for future work.

**Dataset.** The datasets used in our experiments are

relatively small, constrained by the high cost of LLM-based data generation and the impracticality of large-scale manual annotation. Expanding these datasets through scalable self-supervised generation or crowd-sourced pipelines could improve model robustness and generalizability. In addition, although CARPAS was evaluated in finance and public health, the domain coverage remains narrow. Further exploration in areas such as law, education, scientific literature, or sports could better test domain-agnostic applicability and uncover domain-specific challenges in aspect refinement and summarization.

**Method.** We note that the success of the entire framework highly depends on the accuracy of the initial aspect count prediction, which makes the subsequent ABS task brittle. Moreover, when there are hierarchical or nested aspect structures in the documents, #Aspect-RM may no longer be effective. To address this, we plan to explore a more sophisticated agentic approach to identify the correct aspect counts and the aspect structure of each document in future work.

## References


Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: aspect-oriented summarization of news documents. In *ACL (1)*, pages 6494–6506. Association for Computational Linguistics.

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 *EMNLP*, pages 1967–1991. Association for Computational Linguistics.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *EMNLP*, pages 3675–3686. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*. OpenReview.net.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, and 67 others. 2024. Deepseek LLM: scaling open-source language models with longtermism. *CoRR*, abs/2401.02954.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Zhongfen Deng, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Quan Hung Tran, Shuaiqi Liu, Wenting Zhao, Tao Zhang, Yibo Wang, and Philip S. Yu. 2023. Aspect-based meeting transcript summarization: A two-stage approach with weak supervision on sentence classification. In *IEEE Big Data*, pages 636–645. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yichao Feng, Shuai Zhao, Yueqiu Li, Luwei Xiao, Xiaobao Wu, and Anh Tuan Luu. 2025. Aspect-based summarization with self-aspect retrieval enhanced generation. *CoRR*, abs/2504.13054.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *ACL (1)*, pages 6263–6273. Association for Computational Linguistics.

Xiaobo Guo and Soroush Vosoughi. 2024a. Disordered-dabs: A benchmark for dynamic aspect-based summarization in disordered texts. In *EMNLP (Findings)*, pages 416–431. Association for Computational Linguistics.

Xiaobo Guo and Soroush Vosoughi. 2024b. MODABS: multi-objective learning for dynamic aspect-based summarization. In *ACL (Findings)*, pages 2814–2827. Association for Computational Linguistics.

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

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. Ctrlsum: Towards generic controllable text summarization. In *EMNLP*, pages 5879–5915. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *ICLR*. OpenReview.net.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 79 others. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Florian Kunneman, Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2018. Aspect-based summarization of pros and cons in unstructured product reviews. In *COLING*, pages 2219–2229. Association for Computational Linguistics.

Haoyuan Li, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2023. Aspect-aware unsupervised extractive opinion summarization. In *ACL (Findings)*, pages 12662–12678. 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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*.

Mounica Maddela, Mayank Kulkarni, and Daniel Preotiuc-Pietro. 2022. Entsum: A data set for entity-centric extractive summarization. In *ACL (1)*, pages 3355–3366. Association for Computational Linguistics.

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 *ACL/IJCNLP (2)*, pages 1080–1089. Association for Computational Linguistics.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Ankan Mullick, Sombit Bose, Rounak Saha, Ayan Kumar Bhowmick, Aditya Vempaty, Pawan Goyal, Niloy Ganguly, Prasenjit Dey, and Ravi Kokku. 2024. Leveraging the power of llms: A fine-tuning approach for high-quality aspect-based summarization. *CoRR*, abs/2408.02584.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *CoRR*, abs/2503.14023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024. Aclsum: A new dataset for aspect-based summarization of scientific publications. In *NAACL-HLT*, pages 6660–6675. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric P. Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *EMNLP (1)*, pages 6301–6309. Association for Computational Linguistics.

Yong-En Tian, Yu-Chien Tang, Kuang-Da Wang, An-Zi Yen, and Wen-Chih Peng. 2025. Template-based financial report generation in agentic and decomposed information retrieval. In *SIGIR*, pages 2706–2710. ACM.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance. In *EMNLP*, pages 671–686. Association for Computational Linguistics.

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

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *ICLR*. OpenReview.net.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Fangwei Zhu, Shangqing Tu, Jiaxin Shi, Juanzi Li, Lei Hou, and Tong Cui. 2021. TWAG: A topic-guided wikipedia abstract generator. In *ACL/IJCNLP (1)*, pages 4623–4635. Association for Computational Linguistics.

# A Experimental Details

## A.1 Computing Infrastructure

Our experimental setup consisted of a machine running Ubuntu 20.04.6 LTS (Linux kernel 5.15) with an AMD Ryzen Threadripper 3960X CPU, 251 GB of RAM, and two NVIDIA GeForce RTX 3090 GPUs. Key software versions include Python 3.12.2, CUDA 12.2, and PyTorch 2.2.0.

## A.2 Fine-Tuning

We fine-tune the pre-trained embedding model `Qwen/Qwen3-Embedding-0.6B` to perform regression on the number of aspects associated with a given document. To adapt the model for this task, we append a lightweight regression head on top of the document embedding derived from the base model. The document representation is obtained by applying mean pooling over the final hidden states of the encoder. This pooled embedding is passed through a two-layer feed-forward regression head composed of a linear transformation that reduces the hidden size by half, followed by a ReLU activation, a dropout layer with a rate of 0.2, and a final linear projection to a single scalar output. To enable efficient parameter tuning, we adopt LoRA for fine-tuning. The configuration uses a rank of 16, a scaling factor (`lora_alpha`) of 32, and a dropout rate of 0.1. Training is conducted using the AdamW optimizer with a learning rate of $2 \times 10^{-5}$.

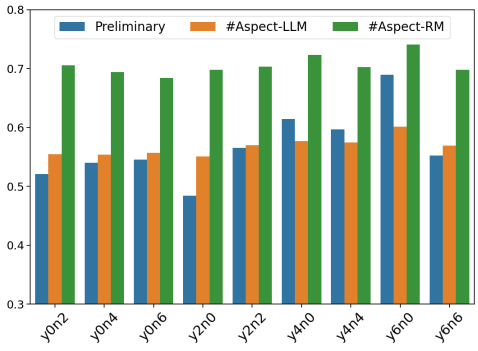



Figure 7: Average BERTScore results on COVID-19-PC across different provided aspect settings.

## A.3 Provided Aspect Settings

In our experiments, we simulate various scenarios by providing the model with a mix of correct and incorrect aspects, denoted as $(y, n)$, where $y$ is the number of correct aspects and $n$ is the number of incorrect aspects. It is important to note a key constraint related to our dataset design: the ground-truth number of aspects in any given document ranges from 4 to 8. Consequently, the number of correct aspects ($y$) provided in any experimental setting is inherently limited by the actual number of ground-truth aspects available in the given document. For instance, consider a document with a ground-truth aspect count of 4. If this document is used in an experiment with the setting $y6n6$, the model will be provided with only 4 correct aspects (since that is the maximum available) along with 6 incorrect aspects. This ensures that the provided correct aspects are genuinely present in the source document.

# B Full Results

## B.1 Quantitative Results

Table 5, 6, and 7 present the complete evaluation metrics, including BERTScore, ROUGE-1, ROUGE-2, and ROUGE-L, for the three datasets. As previously discussed, the #Aspect-RM outperforms both the Preliminary Experiments and the #Aspect-LLM. Furthermore, GPT-based models achieve their best performance when paired with the Self-Refine strategy.

The superior performance of the #Aspect-RM is evident across both COVID-19-PC and RW-ECT. As shown in Figure 9 and 10, it records the lowest

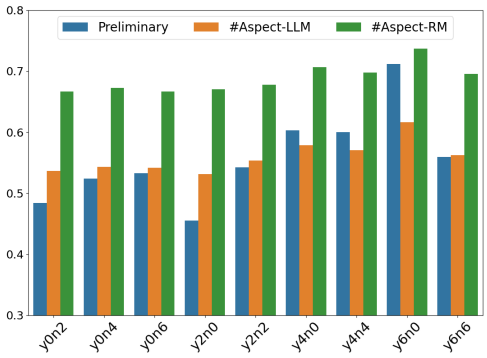



Figure 8: Average BERTScore results on RW-ECT across different provided aspect settings.

#AbsAspDiff, indicating the highest accuracy in aspect count prediction. Consequently, this precision leads to consistently high-quality summaries, with its strong BERTScore performance holding true across all provided aspect settings (Figure 7 and 8).

### B.2 Additional Models

We additionally run experiments with two other reasoning models, o3-mini and Gemini-2.5-Flash-Lite, on the three datasets. As shown in Tables 8, 9, and 10, #Aspect-RM continues to achieve the highest scores, and o3-mini demonstrates the best performance under the Self-Refine strategy.

## C Default Aspect Sets, Dataset & Prompts

We illustrate the dataset construction workflow in Figure 11. After generating the articles, we apply a deduplication process to ensure content diversity and eliminate redundancy. Redundant samples are identified based on the cosine similarity between paired article embeddings computed by `BGE-M3` (threshold = 0.85), as well as Jaccard similarity (0.45) and BLEU score (0.3). Furthermore, we provide the default aspect example in Table 12, and the prompt template and data example in Figure 12, 13, 14, and 15.

| | Prompt | Gemma-3-12B | | | | Gemma-3-27B | | | | GPT-4o-mini | | | | GPT-4o | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L |
| Preliminary | Direct | 56.1 | 39.9 | 26.2 | 32.6 | 62.6 | 42.4 | 25.3 | 33.6 | 50.9 | 31.3 | 16.3 | 23.8 | 62.7 | 40.7 | 22.7 | 31.7 |
| | CoT | 58.1 | 39.8 | 25.9 | 32.8 | 63.5 | 42.1 | 25.7 | 34.3 | 45.3 | 26.6 | 14.5 | 21.0 | 61.6 | 35.9 | 18.9 | 28.3 |
| | CoT-SC | 63.6 | 43.3 | 26.9 | 34.1 | 63.1 | 39.8 | 22.1 | 31.0 | 48.5 | 29.8 | 16.7 | 23.4 | 63.7 | 39.2 | 21.4 | 30.6 |
| | Self-Refine | 56.6 | 39.7 | 25.3 | 32.2 | 51.7 | 35.2 | 21.4 | 28.3 | 53.2 | 33.2 | 18.0 | 25.7 | 56.6 | 35.5 | 18.1 | 26.6 |
| #Aspect-LLM | Direct | 53.5 | 37.7 | 24.5 | 30.7 | 61.6 | 40.7 | 23.5 | 32.0 | 50.9 | 30.5 | 15.5 | 23.0 | 56.3 | 35.9 | 19.5 | 27.6 |
| | CoT | 59.5 | 40.1 | 25.4 | 32.8 | 62.8 | 40.4 | 23.8 | 32.6 | 52.3 | 30.0 | 16.0 | 23.5 | 55.8 | 31.7 | 16.4 | 25.1 |
| | CoT-SC | 59.2 | 40.0 | 24.4 | 31.5 | 60.7 | 37.9 | 20.9 | 29.6 | 56.1 | 33.4 | 18.1 | 26.0 | 60.8 | 37.1 | 19.7 | 28.6 |
| | Self-Refine | 56.1 | 39.2 | 25.2 | 32.0 | 53.2 | 36.3 | 21.9 | 29.0 | 51.7 | 33.8 | 19.3 | 26.6 | 55.3 | 37.3 | 20.9 | 28.7 |
| #Aspect-RM | Direct | 70.7 | **48.4** | **30.4** | **38.6** | 72.9 | **47.5** | 26.6 | 36.7 | 59.7 | 35.5 | 17.6 | 26.7 | 71.5 | 44.3 | 23.6 | 34.0 |
| | CoT | 71.2 | 47.2 | 29.3 | 38.3 | **73.5** | 47.1 | **27.3** | **37.7** | 61.4 | 34.8 | 18.3 | 27.3 | 69.3 | 38.6 | 19.8 | 30.3 |
| | CoT-SC | **72.0** | 46.8 | 26.8 | 35.7 | 71.9 | 44.1 | 23.1 | 33.6 | 69.9 | 40.5 | 21.1 | 31.2 | 71.0 | 42.1 | 21.8 | 32.3 |
| | Self-Refine | 65.2 | 43.1 | 25.7 | 34.2 | 70.6 | 47.2 | 26.8 | 37.0 | **70.6** | **44.0** | **23.4** | **33.5** | **71.9** | **45.9** | **23.9** | **34.4** |

Table 5: Results (×100%) on ECT, with BERT standing for BERTScore, R-1 standing for ROUGE-1, R-2 standing for ROUGE-2, and R-L standing for ROUGE-L. The highest results for each metric are in boldface, while the second-best are underlined.

| | Prompt | Gemma-3-12B | | | | Gemma-3-27B | | | | GPT-4o-mini | | | | GPT-4o | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L | BERT | R-1 | R-2 | R-L |
| Preliminary | Direct | 56.9 | 41.2 | 26.4 | 33.8 | 64.6 | 44.4 | 25.7 | 35.4 | 47.9 | 28.2 | 13.0 | 20.9 | 57.6 | 35.7 | 18.4 | 27.8 |
| | CoT | 55.2 | 38.5 | 24.2 | 31.8 | 64.2 | 43.2 | 25.5 | 35.3 | 40.9 | 22.8 | 11.0 | 17.6 | 60.0 | 34.1 | 17.0 | 27.0 |
| | CoT-SC | 66.2 | 47.5 | 29.7 | 38.1 | 65.2 | 43.4 | 24.5 | 34.7 | 46.2 | 27.8 | 14.2 | 21.4 | 62.2 | 38.6 | 20.4 | 30.4 |
| | Self-Refine | 60.2 | 43.0 | 26.9 | 35.2 | 52.2 | 36.6 | 21.8 | 29.6 | 52.6 | 34.2 | 18.3 | 26.7 | 56.0 | 36.9 | 19.4 | 28.9 |
| #Aspect-LLM | Direct | 54.2 | 39.0 | 24.8 | 32.2 | 61.0 | 41.6 | 23.9 | 33.2 | 55.1 | 32.1 | 14.4 | 23.6 | 56.2 | 34.3 | 17.1 | 26.5 |
| | CoT | 58.8 | 41.0 | 25.8 | 34.0 | 61.4 | 41.3 | 24.3 | 33.6 | 53.4 | 29.1 | 13.7 | 22.4 | 55.8 | 31.1 | 15.1 | 24.5 |
| | CoT-SC | 55.6 | 39.3 | 24.2 | 31.7 | 59.9 | 39.3 | 21.5 | 31.2 | 55.3 | 31.9 | 15.8 | 24.5 | 56.1 | 34.1 | 17.6 | 26.8 |
| | Self-Refine | 53.9 | 37.8 | 23.3 | 30.8 | 60.4 | 42.3 | 24.9 | 34.3 | 54.0 | 33.9 | 17.6 | 26.3 | 57.0 | 36.2 | 18.3 | 28.1 |
| #Aspect-RM | Direct | 69.7 | 49.1 | 30.1 | 39.8 | **74.4** | 49.4 | 27.5 | 38.9 | 64.2 | 36.4 | 15.6 | 26.5 | 71.0 | 42.4 | 20.4 | 32.4 |
| | CoT | 70.1 | 47.9 | 29.5 | 39.4 | 74.2 | 48.9 | 27.9 | 39.4 | 61.6 | 33.3 | 15.3 | 25.4 | 69.0 | 38.2 | 18.4 | 30.1 |
| | CoT-SC | **75.9** | **52.3** | **30.9** | **41.1** | 74.0 | 47.6 | 25.2 | 37.2 | 68.7 | 38.8 | 18.5 | 29.5 | 72.4 | 43.7 | 22.0 | 33.7 |
| | Self-Refine | 66.9 | 46.8 | 28.5 | 38.1 | 72.4 | **50.5** | **29.1** | **40.3** | **71.2** | **44.3** | **22.4** | **34.1** | **73.1** | **46.6** | **23.3** | **35.9** |

Table 6: Results (×100%) on COVID-19-PC.

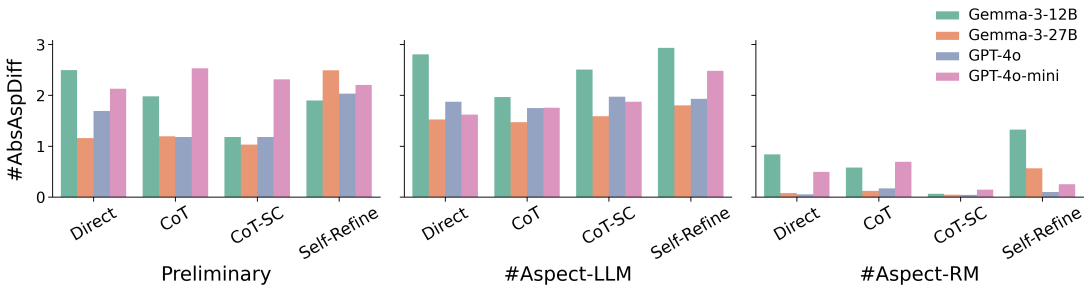



Figure 9: Aspect count error (#AbsAspDiff) across different methods and prompting strategies on COVID-19-PC.

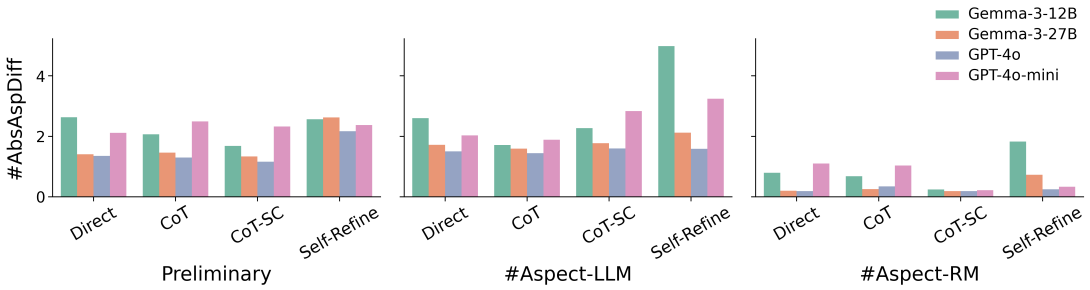



Figure 10: Aspect count error (#AbsAspDiff) across different methods and prompting strategies on RW-ECT.

| | Prompt | Gemma-3-12B | | | | Gemma-3-27B | | | | GPT-4o-mini | | | | GPT-4o | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **BERT** | **R-1** | **R-2** | **R-L** | **BERT** | **R-1** | **R-2** | **R-L** | **BERT** | **R-1** | **R-2** | **R-L** | **BERT** | **R-1** | **R-2** | **R-L** |
| Preliminary | Direct | 54.9 | 37.0 | 22.8 | 29.1 | 62.0 | 39.3 | 21.8 | 30.1 | 48.3 | 27.2 | 12.6 | 19.8 | 61.0 | 36.9 | 19.7 | 27.9 |
| | CoT | 54.6 | 36.4 | 22.7 | 29.5 | 62.3 | 39.5 | 23.2 | 31.6 | 43.0 | 24.1 | 12.2 | 18.3 | 59.8 | 33.2 | 17.0 | 25.6 |
| | CoT-SC | 61.2 | 39.9 | 23.8 | 30.6 | 62.1 | 37.9 | 20.6 | 29.0 | 46.8 | 26.9 | 13.7 | 20.2 | 62.5 | 37.0 | 19.4 | 28.2 |
| | Self-Refine | 55.4 | 37.1 | 22.6 | 29.2 | 51.4 | 34.1 | 20.0 | 26.6 | 51.4 | 32.2 | 16.9 | 24.0 | 54.7 | 34.7 | 18.4 | 25.9 |
| #Aspect-LLM | Direct | 56.0 | 36.9 | 22.3 | 29.1 | 59.0 | 36.8 | 20.0 | 27.9 | 51.1 | 28.6 | 13.3 | 20.7 | 58.7 | 34.7 | 17.6 | 25.7 |
| | CoT | 60.5 | 39.0 | 23.7 | 31.1 | 60.4 | 37.4 | 21.4 | 29.7 | 52.9 | 28.6 | 14.0 | 21.6 | 58.2 | 31.3 | 15.5 | 24.1 |
| | CoT-SC | 56.3 | 35.1 | 19.7 | 26.6 | 57.7 | 34.0 | 17.8 | 25.7 | 51.6 | 28.6 | 14.2 | 21.6 | 58.3 | 33.6 | 17.0 | 25.3 |
| | Self-Refine | 45.4 | 28.2 | 16.1 | 21.9 | 58.4 | 37.9 | 21.6 | 29.3 | 51.0 | 30.0 | 15.2 | 22.5 | 59.1 | 35.9 | 18.0 | 26.2 |
| #Aspect-RM | Direct | 68.9 | **44.6** | **26.2** | **34.7** | 73.0 | 45.0 | 24.2 | 33.9 | 59.4 | 32.8 | 14.8 | 23.6 | 71.4 | 41.8 | 21.1 | 31.2 |
| | CoT | 68.0 | 43.1 | 25.6 | 34.0 | **73.1** | 44.6 | 25.0 | **34.9** | 59.5 | 31.5 | 15.3 | 23.8 | 68.1 | 36.0 | 17.7 | 27.7 |
| | CoT-SC | **71.5** | 43.6 | 23.7 | 32.6 | 71.8 | 41.9 | 21.5 | 31.4 | 69.2 | 37.4 | 18.0 | 27.8 | 71.2 | 40.1 | 20.0 | 30.1 |
| | Self-Refine | 63.0 | 39.5 | 22.4 | 30.4 | 70.3 | **45.3** | **25.3** | 34.7 | **70.2** | **41.2** | **20.4** | **30.7** | **71.7** | **43.2** | **21.5** | **31.7** |

Table 7: Results (×100%) on RW-ECT.

| | **Prompt** | **Gemini-2.5-Flash-Lite** | | | | | **o3-mini-2025-01-31** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **BERT** | **R-1** | **R-2** | **R-L** | **#AbsAspDiff(↓)** | **BERT** | **R-1** | **R-2** | **R-L** | **#AbsAspDiff(↓)** |
| Preliminary | Direct | 0.630 | 0.452 | 0.275 | 0.356 | 1.963 | 0.565 | 0.332 | 0.145 | 0.238 | 1.539 |
| | CoT | 0.608 | 0.436 | 0.266 | 0.347 | 2.176 | 0.489 | 0.282 | 0.130 | 0.213 | 2.190 |
| | CoT-SC | 0.652 | 0.467 | 0.282 | 0.368 | 1.650 | 0.542 | 0.329 | 0.152 | 0.239 | 1.787 |
| | Self-Refine | 0.503 | 0.363 | 0.227 | 0.293 | 3.957 | 0.554 | 0.351 | 0.169 | 0.258 | 1.971 |
| #Aspect-LLM | Direct | 0.512 | 0.367 | 0.228 | 0.293 | 4.561 | 0.589 | 0.342 | 0.145 | 0.242 | 1.098 |
| | CoT | 0.496 | 0.353 | 0.213 | 0.281 | 4.831 | 0.582 | 0.320 | 0.139 | 0.238 | 1.249 |
| | CoT-SC | 0.504 | 0.356 | 0.211 | 0.280 | 4.246 | 0.604 | 0.345 | 0.148 | 0.247 | 1.087 |
| | Self-Refine | 0.496 | 0.357 | 0.223 | 0.287 | 4.536 | 0.619 | 0.379 | 0.172 | 0.272 | 1.088 |
| #Aspect-RM | Direct | 0.741 | 0.515 | **0.302** | 0.398 | 0.396 | 0.661 | 0.376 | 0.156 | 0.266 | 0.277 |
| | CoT | 0.735 | 0.510 | 0.298 | **0.399** | 0.525 | 0.656 | 0.358 | 0.155 | 0.266 | 0.465 |
| | CoT-SC | **0.750** | **0.516** | 0.293 | 0.396 | **0.307** | 0.680 | 0.385 | 0.163 | 0.275 | 0.257 |
| | Self-Refine | 0.635 | 0.454 | 0.278 | 0.361 | 1.613 | **0.697** | **0.422** | **0.191** | **0.303** | **0.254** |

Table 8: Results of two reasoning models on ECT. BERT refers to BERTScore, R-1 to ROUGE-1, R-2 to ROUGE-2, R-L to ROUGE-L, and #AbsAspDiff is the absolute aspect difference.

| | **Prompt** | **Gemini-2.5-Flash-Lite** | | | | | **o3-mini-2025-01-31** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **BERT** | **R-1** | **R-2** | **R-L** | **#AbsAspDiff(↓)** | **BERT** | **R-1** | **R-2** | **R-L** | **#AbsAspDiff(↓)** |
| Preliminary | Direct | 0.610 | 0.436 | 0.259 | 0.351 | 1.874 | 0.569 | 0.334 | 0.140 | 0.240 | 1.451 |
| | CoT | 0.533 | 0.373 | 0.220 | 0.303 | 2.721 | 0.489 | 0.270 | 0.112 | 0.202 | 1.933 |
| | CoT-SC | 0.586 | 0.421 | 0.252 | 0.341 | 2.325 | 0.556 | 0.337 | 0.148 | 0.245 | 1.519 |
| | Self-Refine | 0.506 | 0.368 | 0.225 | 0.304 | 3.448 | 0.557 | 0.351 | 0.159 | 0.258 | 1.834 |
| #Aspect-LLM | Direct | 0.512 | 0.367 | 0.221 | 0.299 | 3.601 | 0.604 | 0.347 | 0.139 | 0.246 | 0.882 |
| | CoT | 0.472 | 0.336 | 0.198 | 0.274 | 4.456 | 0.588 | 0.316 | 0.126 | 0.234 | 0.974 |
| | CoT-SC | 0.499 | 0.355 | 0.210 | 0.288 | 3.635 | 0.616 | 0.357 | 0.147 | 0.256 | 0.866 |
| | Self-Refine | 0.495 | 0.358 | 0.219 | 0.295 | 3.946 | 0.633 | 0.390 | 0.150 | 0.281 | 0.885 |
| #Aspect-RM | Direct | 0.739 | 0.511 | **0.293** | 0.407 | 0.275 | 0.680 | 0.385 | 0.151 | 0.273 | 0.053 |
| | CoT | 0.708 | 0.485 | 0.274 | 0.388 | 0.578 | 0.665 | 0.356 | 0.140 | 0.262 | 0.159 |
| | CoT-SC | **0.753** | **0.515** | 0.291 | **0.410** | **0.128** | 0.695 | 0.399 | 0.162 | 0.285 | **0.041** |
| | Self-Refine | 0.665 | 0.475 | 0.282 | 0.386 | 1.222 | **0.714** | **0.432** | **0.186** | **0.312** | **0.041** |

Table 9: Results of two reasoning models on COVID-19-PC.

| | Prompt | Gemini-2.5-Flash-Lite | | | | | o3-mini-2025-01-31 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BERT | R-1 | R-2 | R-L | #AbsAspDiff(↓) | BERT | R-1 | R-2 | R-L | #AbsAspDiff(↓) |
| Preliminary | Direct | 0.581 | 0.396 | 0.237 | 0.308 | 3.047 | 0.549 | 0.302 | 0.123 | 0.211 | 1.592 |
| | CoT | 0.588 | 0.409 | 0.247 | 0.316 | 2.232 | 0.464 | 0.253 | 0.111 | 0.187 | 2.245 |
| | CoT-SC | 0.625 | 0.435 | 0.263 | 0.338 | 1.919 | 0.525 | 0.299 | 0.129 | 0.214 | 1.744 |
| | Self-Refine | 0.503 | 0.349 | 0.211 | 0.273 | 3.734 | 0.548 | 0.330 | 0.148 | 0.234 | 1.939 |
| #Aspect-LLM | Direct | 0.478 | 0.327 | 0.196 | 0.255 | 5.013 | 0.556 | 0.297 | 0.116 | 0.207 | 1.303 |
| | CoT | 0.466 | 0.318 | 0.191 | 0.251 | 5.457 | 0.555 | 0.288 | 0.121 | 0.213 | 1.414 |
| | CoT-SC | 0.477 | 0.319 | 0.186 | 0.248 | 4.532 | 0.561 | 0.299 | 0.119 | 0.211 | 1.370 |
| | Self-Refine | 0.475 | 0.325 | 0.197 | 0.258 | 5.144 | 0.583 | 0.332 | 0.139 | 0.231 | 1.303 |
| #Aspect-RM | Direct | 0.728 | 0.481 | 0.271 | 0.366 | 0.424 | 0.657 | 0.348 | 0.133 | 0.243 | 0.205 |
| | CoT | 0.722 | 0.482 | **0.278** | **0.373** | 0.595 | 0.658 | 0.340 | 0.140 | 0.249 | 0.338 |
| | CoT-SC | **0.743** | **0.486** | 0.272 | 0.371 | **0.252** | 0.670 | 0.351 | 0.137 | 0.247 | **0.181** |
| | Self-Refine | 0.673 | 0.458 | 0.271 | 0.356 | 1.175 | **0.690** | **0.391** | **0.167** | **0.274** | **0.181** |

Table 10: Results of two reasoning models on RW-ECT.

| Dataset | Method | Gemini-2.5-Flash | | | Gemma-3-12B | | | Gemma-3-27B | | | GPT-4o-mini | | | GPT-4o | | | O3-mini | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AS | BS(↑) | AAD(↓) | AS | BS(↑) | AAD(↓) | AS | BS(↑) | AAD(↓) | AS | BS(↑) | AAD(↓) | AS | BS(↑) | AAD(↓) | AS | BS(↑) | AAD(↓) |
| ECT | Preliminary | 3.36 | 0.50 | 3.96 | 3.20 | 0.57 | 2.42 | 2.30 | 0.52 | 2.65 | 11.16 | 0.53 | 3.15 | 2.75 | 0.57 | 2.21 | 3.34 | 0.55 | 1.97 |
| | #Aspect-LLM | 3.39 | 0.50 | 4.54 | 3.14 | 0.56 | 2.53 | 2.30 | 0.53 | 2.39 | 3.42 | 0.53 | 2.92 | 3.24 | 0.55 | 2.10 | 3.90 | 0.62 | 1.09 |
| | #Aspect-RM | 3.38 | 0.63 | 1.61 | 5.09 | 0.65 | 1.31 | 3.59 | 0.71 | 0.78 | 7.15 | 0.71 | 0.42 | 4.17 | 0.72 | 0.29 | 3.86 | 0.70 | 0.25 |
| COVID-19-PC | Preliminary | 3.32 | 0.51 | 3.45 | 3.21 | 0.60 | 1.90 | 2.52 | 0.52 | 2.49 | 3.58 | 0.53 | 2.20 | 3.36 | 0.56 | 2.03 | 3.43 | 0.56 | 1.83 |
| | #Aspect-LLM | 3.67 | 0.50 | 3.95 | 5.86 | 0.54 | 2.93 | 3.65 | 0.60 | 1.80 | 5.44 | 0.54 | 2.48 | 4.00 | 0.57 | 1.93 | 3.92 | 0.63 | 0.89 |
| | #Aspect-RM | 3.45 | 0.67 | 1.22 | 6.15 | 0.67 | 1.32 | 3.56 | 0.72 | 0.56 | 5.08 | 0.71 | 0.25 | 3.85 | 0.73 | 0.10 | 3.86 | 0.71 | 0.04 |
| RW-ECT | Preliminary | 3.25 | 0.50 | 3.73 | 3.24 | 0.55 | 2.56 | 2.32 | 0.51 | 2.62 | 3.59 | 0.51 | 2.37 | 3.27 | 0.55 | 2.17 | 3.38 | 0.55 | 1.94 |
| | #Aspect-LLM | 4.01 | 0.48 | 5.14 | 5.03 | 0.45 | 4.98 | 3.39 | 0.58 | 2.11 | 7.10 | 0.51 | 3.24 | 3.95 | 0.59 | 1.59 | 3.95 | 0.58 | 1.30 |
| | #Aspect-RM | 3.53 | 0.67 | 1.18 | 6.53 | 0.63 | 1.82 | 3.43 | 0.70 | 0.73 | 6.91 | 0.70 | 0.33 | 3.90 | 0.72 | 0.24 | 3.85 | 0.69 | 0.18 |

Table 11: Results of Self-Refine on the three dataset, showing average thinking steps (AS), BERTScore (BS), and #AbsAspDiff (AAD).

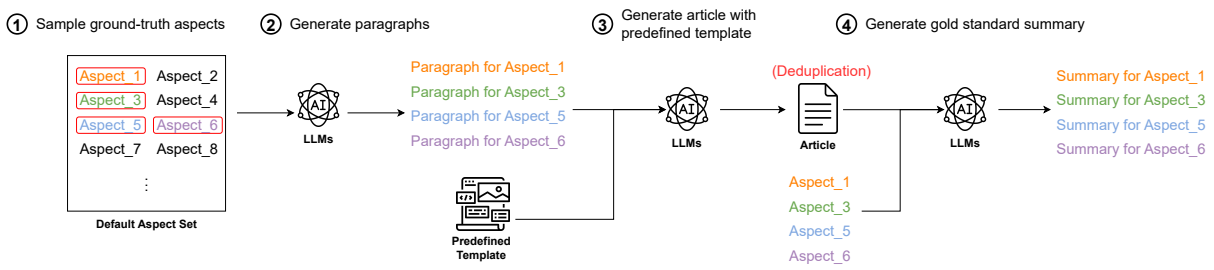



Figure 11: The illustration of synthetic data generation.

| Dataset | Default Aspect Set |
|---|---|
| ECT & RW-ECT | • **Profit and Loss Statement Highlights**<br>– Revenue results, QoQ changes, YoY changes with reasons, revenue results, and guidance<br>– Wafer sales and the breakdown to wafer quantity and ASP<br>– Net margin and Diluted EPS (earnings per share) results<br>• **Segment or Platform Highlights**<br>– Sales by segment or platforms, their respective margin levels, and their respective management comments<br>– Sales guidance, forecast, or trend by segment next quarter or full year<br>• **Fab Construction, Expansion and Capacity**<br>– New fab construction progress for this quarter<br>– Capacity expansion plan on the existing fabs for this quarter |
| COVID-19-PC | • **Epidemic Statistics Overview for the Day**<br>– Total confirmed cases (local and imported), Day-over-Day or Week-over-Week changes with context<br>– Total deaths and recovered cases, breakdown by age or comorbidity where applicable<br>– Testing numbers (e.g. PCR, antigen), positivity rate trends<br>• **Vaccination Progress and Plans**<br>– Daily/weekly vaccination numbers, coverage rates by age group<br>– Updates on booster doses or new vaccine arrivals<br>– Comments on vaccine supply, procurement, and logistics<br>• **Public Health Policy Updates**<br>– Changes to mask mandates, social distancing, gathering limits<br>– Border control measures: entry rules, quarantine requirements<br>– Changes to public health alert levels (e.g., from level 3 to 2) |

Table 12: Example of default aspect set.

**Example of ECT Template (4 Aspects)**

Operator:
Good day and thank you for standing by.
Welcome to the {{COMPANY_NAME}} {{QUARTER}} Earnings Conference Call.
At this time, all participants are in a listen-only mode. After the speakers' presentation, there will be a Q&A session.
Today's call is being recorded.
I would now like to hand the call over to {{IR_NAME}}, {{IR_TITLE}}. Please go ahead.

{{IR_NAME}}, {{IR_TITLE}}:
Thank you, Operator. Good {{TIME_OF_DAY}}, and welcome to {{COMPANY_NAME}}'s earnings call.
Joining me today are {{CEO_NAME}}, CEO; {{CFO_NAME}}, CFO; and {{EXEC_NAME}}, {{EXEC_TITLE}}.
Earlier today, we published our earnings press release and slides, which are available on our IR site.
Please note this call may contain forward-looking statements.
Now, I'll turn it over to {{CEO_NAME}}.

{{CEO_NAME}}, Chief Executive Officer:
Thanks, {{IR_NAME}}. Hello, everyone.

Let's start with a look at the quarter:
This quarter, we achieved significant progress, including {{ASPECT_1_DETAILS}}.
We're seeing strong performance in this area, supported by {{ASPECT_2_DETAILS}}.

These results reflect our strategy in action. I'll now hand it over to {{CFO_NAME}} to discuss the financials.

{{CFO_NAME}}, Chief Financial Officer:
Thanks, {{CEO_NAME}}, and hello, everyone.

Let's take a closer look at the numbers:
{{ASPECT_2_DETAILS}}
{{ASPECT_3_DETAILS}}

Now to {{EXEC_NAME}}, who will cover operational and product updates.

{{EXEC_NAME}}, {{EXEC_TITLE}}:
Thanks, {{CFO_NAME}}.

Let's revisit {{ASPECT_1}} from an innovation standpoint.
We launched {{NEW_PRODUCT}} which contributed to {{IMPACT_OF_PRODUCT}}.
We've made strong headway with {{ASPECT_4_DETAILS}}, and expect meaningful results in {{ASPECT_4_TIMELINE}}.

Let's move to Q&A.

Analyst 1:
On {{ASPECT_2}}, could you expand on?

{{CFO_NAME}}:
Absolutely. We believe {{ASPECT_2_FOLLOWUP}}.

Analyst 2:
Regarding {{ASPECT_4}}, what's the scale and outlook?

{{EXEC_NAME}}:
We're seeing momentum with {{ASPECT_4_IMPACT}}, especially in...

{{CEO_NAME}}:
Thank you all for joining. We're proud of our progress and excited for what's ahead.

Operator:
Thank you. This concludes the call. You may now disconne

Figure 12: Example template of ECT with 4 aspects.

## Synthetic Data Generation Prompt

**Generate Detailed Aspect Descriptions:**

You are a financial analyst helping generate synthetic financial content.
Write a highly detailed and plausible explanation about the following financial aspects: {sample aspects}
Include fictional but realistic figures, trends, technical initiatives, regional commentary, and strategic insights.
Be as specific as possible.

**Generate Document from Aspects:**

You are a financial content writer creating a realistic earnings call transcript (3000–5000 words).

You'll receive a speaker template, a list of financial aspects, and an Aspect Descriptions dictionary with detailed content. All descriptions must be fully paraphrased and naturally integrated into the dialogue — they are the complete source material, not summaries or hints.Write a natural, flowing transcript following the speaker order. Use a professional, executive tone. No bullet points, titles, or structural references. Do not mention "aspects" or "templates." You may assign fictional names for the company, executives, and products.Every description must be fully and smoothly covered in the conversation. Return only the final transcript as plain text.

Template:
{template}

Aspects:
{provided aspects}

Aspect Description
{aspect descriptions}

**Generate Aspect-Based Summary:**

You are a financial analyst tasked with extracting structured insights from a given article.

You will be provided with a list of financial aspects. Your responsibility is to carefully read the article and, for each aspect, write a concise and accurate summary that reflects exactly what the article says about it.

Article:
{article}

Aspects:
{provided aspects}

Return only the final result in this format:

```
{{
  "Aspect Title 1": "summary...",
  "Aspect Title 2": "summary...",
  ...
}}
```

Figure 13: Example prompts on synthetic data generation on ECT.

## ECT

**Article:**
...
Emily Carter, Chief Financial Officer:
Thanks, Alan, and hello, everyone.
Let's take a closer look at the numbers:
As Alan mentioned, we are projecting total revenue for Q3 2024 to be in the range of $125 million to $130 million. That's a growth of 3% to 7% compared to Q3 2023...

Moving on to gross margin, we are projecting a range of 68% to 70% for Q3, slightly lower than the 71.5% we reported in Q3 2023. This is primarily due to increased cloud infrastructure costs and the impact of promotional pricing in key markets... For Q4, we anticipate a gross margin in the range of 48.0% to 49.0%.
...

Question 3:
Can you provide some color on the wafer ASP trends you are seeing in the market and how that is impacting your costs?

Emily Carter:
Okay, Michael, I can provide some insight on that. Overall wafer ASP in Q3 2024 experienced a moderate increase of 2.5% QoQ, landing at an average of $1,250 per wafer equivalent... In terms of node specifics, ASPs for 3nm wafers saw the most significant increase, jumping by 8% QoQ to an average of $18,000 per wafer. At the 5nm node, ASPs remained relatively stable, showing a modest increase of 1% QoQ to approximately $12,500 per wafer. However, for 7nm wafers, we saw a decline of 2% QoQ to an average of $8,000 per wafer...

**Aspect Summary**:

Overall wafer ASP comments, and the by node or segment wafer ASP comments: Overall wafer ASP in Q3 2024 experienced a moderate increase of 2.5% QoQ, averaging $1,250 per wafer equivalent, slightly below the projected 3%. This was due to demand mix and pricing pressure in mature nodes. 3nm wafers saw an 8% QoQ increase to $18,000, 5nm wafers increased modestly by 1% QoQ to $12,500, 7nm wafers declined by 2% QoQ to $8,000, and mature nodes (40nm-90nm) declined slightly by 1% QoQ to $800.

Gross margin results: Q3 2024 gross margin was 48.5%, a 0.8 percentage point decrease QoQ and a 1.2 percentage point increase YoY. Compared to the Q2 guidance of 49.0%, there was a variance of -0.5 percentage points due to unforeseen raw material cost inflation and product mix shift. The QoQ decrease was driven by raw material cost inflation, increased promotional activity in North America, and foreign exchange headwinds. The YoY increase was driven by supply chain optimization, product innovation and premiumization, and manufacturing efficiency improvements.

CapEx guidance: The full-year 2024 CapEx guidance was revised upwards to $750 million, a 15% increase from the initial guidance of $650 million. This revision primarily reflects accelerated investment in AI infrastructure, specifically GPU capacity and high-bandwidth networking, accounting for roughly 45% of the total increase, or $337.5 million. This includes the acquisition and deployment of advanced GPU clusters, primarily NVIDIA H100 and AMD MI300 series, and high-performance networking equipment.

P&L (revenue and gross margin) guidance: The company is projecting total revenue for Q3 2024 to be in the range of $125 million to $130 million, representing a growth of 3% to 7% compared to Q3 2023. Gross margin is projected to be in the range of 68% to 70% for Q3, slightly lower than the 71.5% reported in Q3 2023. For Q4, they anticipate a gross margin in the range of 48.0% to 49.0%.

Figure 14: Example ECT dataset.

**COVID-19-PC**

**Article:**

...
Dr. Humphrey:
Thank you, Eleanor. Good afternoon, everyone.

Let's start with today's situation: As of this morning, the total number of confirmed COVID-19 cases in the country stands at 12,475,321. This figure includes both locally transmitted cases and imported cases detected at our borders. Over the past week (October 27th to November 2nd), we have observed a 7.8% increase in confirmed cases compared to the previous week (October 20th to October 26th)... Yesterday we recorded 21,542 new cases, while the day before that the number was 19,876.
...
Dr. Sharma:
Now, concerning new variants: Over the past few weeks, our genomic surveillance efforts have identified two variants of concern that require our immediate and focused attention: Variant Eta (η) and Variant Iota (ι). Variant Eta, first detected in late March, appears to be approximately 30-40% more transmissible than the previously dominant Delta variant... Variant Iota, initially identified in the Pacific Northwest, possesses mutations that may reduce the effectiveness of certain monoclonal antibody treatments and, potentially, the durability of vaccine-induced immunity.
...
Mr. Davies:
We are pleased to announce that we have recently received a substantial shipment of the Moderna and Pfizer-BioNTech bivalent boosters, totaling 12 million doses... The Novavax COVID-19 vaccine (Nuvaxovid) is now available as a booster option for adults aged 18 years and older... individuals are now eligible for a bivalent booster at least two months after completing a primary series or receiving their last COVID-19 vaccine dose.

Finally, since the rollout of the COVID-19 vaccines, we have been meticulously monitoring adverse events through the Vaccine Adverse Event Reporting System (VAERS) and the Vaccine Safety Datalink (VSD). The COVID-19 vaccines authorized for use continue to demonstrate a very favorable safety profile... Common side effects include pain or swelling at the injection site, fatigue, headache, muscle aches, chills, and fever... We have also been closely tracking rare but more serious adverse events like myocarditis and pericarditis.

**Aspect Summary**:

Clarification of vaccine side effects or variant transmission: Vaccines show a favorable safety profile with mainly mild side effects; rare events like myocarditis are monitored. mRNA vaccines are the preferred option. Variant Eta is 30-40% more transmissible than Delta, and Iota poses a risk of breakthrough infections.

Total confirmed cases (local and imported), Day-over-Day or Week-over-Week changes with context: Total confirmed cases reached 12,475,321. A 7.8% week-over-week increase in cases signals a slight acceleration in transmission.

Updates on booster doses or new vaccine arrivals: Booster supply was increased with 12 million bivalent doses and the addition of Novavax. Eligibility is now 2 months after the last dose, with uptake reaching 28% among eligible adults.

New variants detected and their potential impact: Two new variants of concern, Eta and Iota, have been detected. Eta is 30-40% more transmissible than Delta, while Iota may reduce treatment efficacy and cause breakthrough infections. Response measures include updated mask advisories and an FDA review of treatments.

Figure 15: Example COVID-19-PC dataset.

**Direct Prompt**

You are given a financial article and a list of section titles ("aspects"). Your task is to revise this list to better reflect the article's key themes and content.

For each aspect, decide whether to Retain (if it's well-phrased, specific, and clearly supported), Modify (if it's relevant but needs rewording or more focus), or Delete (if it's irrelevant, redundant, or unsupported).

Identify any important but missing aspects. For each one, provide a clear title and a brief summary of what the article says about it.

For each finalized aspect (retained, modified, or newly added), write a concise summary of what the article says about that topic.

Make sure all aspects are clear, specific, and non-overlapping. Do not include deleted aspects in the final output. Summaries must be factual and based on the article.
{predict aspect number} // Only exists on #Aspect-LLM and #Aspect-RM

Only return your result in the following JSON format:
{{
"Aspect Title 1": "summary for aspect 1...",
"Aspect Title 2": "summary for aspect 2...",
...
}}

Article:
{article}

Provided Aspects:
{provided aspects}

**CoT Prompt**

You are a financial analyst assistant tasked with revising a list of section titles (called "aspects") based on the article content below. Use a chain-of-thought approach: think **step-by-step** to evaluate whether each aspect is relevant and what summary should accompany it.

(Same as Direct Prompt)

.
.
.

Figure 16: Example prompts of Direct and CoT strategies on ECT.

**CoT-SC Final Answer Selection Prompt**

You are given multiple structured outputs from different agents who analyzed a financial article based on specific aspects. Your task is to merge and finalize the aspect-based summaries.

Include only aspects that are clearly supported by the article.
Merge or generalize similar or overly specific aspects into broader, meaningful themes.
Choose the most consistent and accurate summary for each aspect, or write a better one if needed.
{predict aspect number} // Only exists on #Aspect-LLM and #Aspect-RM

Input:
{multiple CoT outputs}

Return only the final result in this format:
{{
"Aspect Title 1": "summary...",
"Aspect Title 2": "summary...",
...
}}

**Self-Refine Prompt**

**Check Coverage:**
You are given an article and a list of aspect titles. Your task is to decide whether these aspects better reflect the key themes or main topics.
{predict aspect number} // Only exists on #Aspect-LLM and #Aspect-RM

If yes, answer only: YES
If not, answer only: NO

Article:
{article}

Aspects:
{provided aspects}

**Refine Aspects:**
You are given a financial article and a list of initial section titles ("aspects"). Improve the list.
For each aspect, say whether to Retain (clear, specific, well-supported), Modify (relevant but needs rewording or focus), or Delete (irrelevant or unsupported).
Add any missing but important aspects by giving a clear title and briefly summarizing what the article says.
{predict aspect number} // Only exists on #Aspect-LLM and #Aspect-RM

Make sure all aspects are clear, specific, non-overlapping, and reflect the main themes of the article.

Article:
{article}

Aspects:
{provided aspects}

**Summarize:**
You are given an article and a list of finalized aspects. Write a concise summary for each aspect based on the article.

Article:
{article}

Aspects:
{provided aspects}

Return only the final result in this format:
{{
"Aspect Title 1": "summary...",
"Aspect Title 2": "summary...",
...
}}

Figure 17: Example prompts of CoT-SC final selection and Self-Refine strategies on ECT.