Aspect–Sentiment Quad Prediction with Distilled Large Language Models

Filippos Ventirozos[1,2], Peter Appleby[2], Matthew Shardlow[1]
[1]Manchester Metropolitan University,
[2]Autotrader Research Group, Autotrader UK
f.ventirozos@mmu.ac.uk

Abstract

Aspect-based sentiment analysis offers detailed insights by pinpointing specific product aspects in a text that are associated with sentiments. This study explores it through the prediction of quadruples, comprising aspect, category, opinion, and polarity. We evaluated in-context learning strategies using recently released distilled large language models, ranging from zero to full-dataset demonstrations. Our findings reveal that the performance of these models now positions them between the current state-of-the-art and significantly higher than their earlier generations. Additionally, we experimented with various chain-of-thought prompts, examining sequences such as aspect to category to sentiment in different orders. Our results[1] indicate that the optimal sequence differs from previous assumptions. Additionally, we found that for quadruple prediction, few-shot demonstrations alone yield better performance than chain-of-thought prompting.

1 Introduction

Tracking customer satisfaction is a pivotal element for organisations striving to enhance their products and services. In a digital world, written feedback can be analysed for positive or negative statements using the NLP techniques of sentiment analysis, a key subfield of text classification. However, this approach is often applied to a full-text review, lacking granularity. A single text can contain multiple opinions aimed at different aspects of a product or a service. This results in ambiguous classifications and limits actionable insights. To overcome this limitation, aspect-based sentiment analysis (ABSA) methodologies (Zhang et al., 2023b) have been investigated to accurately capture the nuanced sentiments in customer feedback.

[1]Code available at: github.com/FilipposVentirozos/Aspect-Sentiment-Quad-Prediction-with-Distilled-Large-Language-Models

In our study, we investigated the compound ABSA task of aspect sentiment quad prediction (ASQP) using large language models (LLMs). This task is the most comprehensive under ABSA, as it maximises information extraction from the text (Zhang et al., 2023b). For a given sentence, ASQP considers the aspects, categories, opinions, and polarities, as shown in Figure 1. This example demonstrates two quadruples from the same text. Firstly, the review found the pizza (Aspect), a type of FOOD (Category), to be delicious (Opinion), representing positive sentiment (Polarity). Secondly, the review found the service (Aspect and Category) to be terrible (Opinion), representing negative polarity.

Quadruple Example

Input: *The pizza is delicious but the service is terrible.*

Output:

Aspect	Category	Polarity	Opinion
pizza	FOOD	POS	delicious
service	SERVICE	NEG	terrible

Figure 1: A quadruple parsing example with two quadruples extracted from one review.

In this paper, we review in-context learning (ICL) techniques for ASQP and apply these techniques to recently released distilled LLMs, specifically GPT-4o (OpenAI et al., 2024), Gemini 2.0 Flash (Gemini et al., 2024) and Qwen3-30-A3B (Yang et al., 2025). ICL involves providing models with a set of examples within the input context to enable the model to make predictions based on these examples, effectively allowing the model to produce accurate responses to unseen tasks without further fine-tuning. We experimented with vary-

Proceedings of Recent Advances in Natural Language Processing,pages 1309–1319
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-152

ing the provided set of examples and analysed the models' performance.

Furthermore, we investigated chain-of-thought (CoT) reasoning, a type of ICL which requires intermediate steps of analysis to be generated. CoT allows additional interpretability of the final results. Evidence also shows that reasoning approaches outperform single-turn answer generation (Wei et al., 2022). This method facilitates a more structured reasoning process, leading to more precise sentiment analysis. Previous research suggested (Fei et al., 2023; Wang and Luo, 2023) retrieving aspects first, followed by opinions and then polarities. In our experiments, we adapted this for the ASQP domain and questioned CoT assumptions by experimenting with different possible CoT chains, such as determining the sentiment before identifying the aspect or vice versa.

We compared these approaches against state-of-the-art methodologies from the literature on widely used ASQP benchmark datasets. Our experiments demonstrate the efficacy of these approaches, highlighting the potential for improved performance in ASQP tasks. The primary contributions of this study are enumerated below:

1. We evaluated ICL using an expanded context window containing a substantially larger number of few-shot examples for ASQP and two additional ABSA tasks.
2. We assessed a diverse set of CoT agents to determine their effectiveness in ASQP as well as in broader ABSA applications.
3. We analysed and compared the performance and output characteristics of GPT-4o, Gemini 2.0 Flash and Qwen3-30B-A3B.

2 Related Work

2.1 Aspect Sentiment Quad Prediction

Aspect-sentiment quad prediction (ASQP) indicates the extraction of sentiment quadruples from text, as shown in Figure 1.

The seminal study by Wan et al. (2020) addressed a significant gap in ABSA tasks. Traditionally, these tasks extracted polarities from either the aspects or the categories, but not both simultaneously. The researchers introduced a new framework called target-aspect-sentiment joint detection. This framework included all the critical elements except the opinion, thus forming a triplet. Their work paved the way for advancements in ASQP.

Building on this foundational work, recent literature has predominantly experimented with fine-tuning encoder-decoder models for ASQP and focused on data augmentation techniques. For instance, Zhang et al. (2021b)[GAS] introduced an encoder-decoder approach using the T5 model (Raffel et al., 2020) to fine-tune the model on inferring aspects, categories, and polarities. Following this, Zhang et al. (2021a)[Para] demonstrated that transforming sentiment quadruples into natural language sentences yielded better results. Under the same output context, Hu et al. (2022)[DLO] sought to refine the generated quadruples by selecting the appropriate order of elements based on entropy calculated from the decoder component.

Peper and Wang (2022)[GEN-S-N] managed to obtain better results on quadruple prediction with implicit aspects and opinions by employing contrastive learning and engineering the target format output for more natural generation.

Recent studies utilising encoder-decoders have increasingly focused on generating additional instances to address data imbalances in ASQP. Yu et al. (2023)[DAST] employed the masked language modelling technique and synonym replacement to augment the data, refining the examples through an iterative methodology. Then Wang et al. (2023)[GenDA] proposed a generative data augmentation technique to optimise encoder-decoder transformer fine-tuning. They created new training instances by swapping aspects and opinions within the same categories, followed by filtering and balancing the dataset.

Gou et al. (2023)[MvP] found that generating quadruple elements in different orders and subsequently selecting based on entropy yields better results. In contrast, we experiment with a multi-hop setting, querying the LLM in each turn to retrieve individual elements and subsequently forming the quadruples at the end in a consistent element order.

2.2 Decoders & In-Context Learning

Decoder-type transformers are popular for their ease of training and scaling (Brown et al., 2020). Major companies have widely adopted LLM decoders via easy-to-use APIs, eliminating extensive infrastructure needs. The success of models like OpenAI's GPT-4 (OpenAI et al., 2024) and Google's Gemini (Gemini et al., 2024) highlights their effectiveness in diverse NLP tasks.

These decoders excel in zero or few-shot learn-

ing scenarios, commonly using ICL. ICL adapts language model output based on the input prompt, using examples for demonstration to perform inference or prediction tasks (Dong et al., 2023). Unlike explicit fine-tuning, ICL can be seen as meta-optimisation. Dai et al. (2023) note that ICL and fine-tuning share similarities in altering attention weights when LLMs tackle NLP tasks.

ICL has been explored for the ASQP task. Xu et al. (2023) were the first to use the ChatGPT model (spec. GPT-3.5-turbo) for this task, devising a prompt template and experimenting with various ASQP demonstrations. Similarly, Zhang et al. (2023a) conducted a systematic analysis comparing ChatGPT with T5, an encoder-decoder fine-tuning approach. They found that ChatGPT sets a robust baseline, requiring T5 to use five to ten times more data to achieve comparable performance. However, with more training samples or demonstrations, T5 can steadily achieve better results, while ChatGPT's performance varied per task, generally favouring ABSA tasks. More recently, Bai et al. (2024) investigated the application of LLMs to a range of ABSA tasks using ICL examples. In comparison, we focus specifically on the ASQP task, evaluate alternative prompting techniques, and base our experiments on distilled LLMs, which are easier to deploy and computationally more efficient.

2.3 Chain of Thought

One subtype of in-context learning is the CoT approach. In CoT, the input prompt is designed to guide the model in breaking down the task into a series of steps. This method has demonstrated improvements across various NLP tasks (Wei et al., 2022). In the following referenced literature and this current study, we specifically examine multi-turn CoT interactions. This technique involves iterative dialogues between the model and the user. By simulating such conversations, we can effectively guide the model to generate high-quality outputs.

Fei et al. (2023) introduced a three-hop reasoning framework for sentiment analysis. This involved a CoT prompting, where they would initially ask the LLM to retrieve the aspects from a text, then append the answer of that in the following prompt query, which asked the opinions on those aspects. Lastly, the above generations would be appended to the final query, which would ask the sentiment polarity. They followed this CoT chain: 1) aspects 2) opinion 3) polarity. Likewise, Wang and Luo (2023) evaluated using CoT for the sentiment analysis text classification. In addition, they analysed the LLM ability of role-playing. And then they chained the CoT following the same CoT chain.

In our study, we delve deeper into the most challenging task of ABSA, the ASQP. While previous approaches have applied CoT techniques to sentiment analysis, specifically text classification, there are none in ASQP, which necessitates a more fine-grained approach. Additionally, we conduct a statistical analysis to determine which CoT sequence yields more accurate results. Specifically, we compare the effectiveness of different CoT chains, such as *sentiments* $\rightarrow$ *aspects* $\rightarrow$ *categories* versus *aspects* $\rightarrow$ *sentiments* $\rightarrow$ *categories*, to identify the optimal sequence for improving ASQP performance.

3 Methodology

3.1 Problem Statement

Our problem statement is identical to that of previous work on ASQP. Given a text, specifically a sentence, we aim to extract zero to multiple quadruples from the text. Each quadruple should contain the elements: aspect, category, opinion, and polarity, as illustrated in Figure 1, and we consider it correct when the generation retains this specified order. While the order of the quadruples may vary, the order of the four elements within each quadruple should remain consistent. Such as:

$$Q = \{(a_i, c_i, p_i, o_i)\}_{i=1}^{n} \quad (1)$$

where Q represents the set of quadruples, a_i is the aspect, c_i is the category, p_i is the polarity, and o_i is the opinion for the i-th quadruple, and n is the number of quadruples extracted from the text. In the following subsections, we describe the different methods utilised in our experiments.

3.2 In-Context Learning

We utilised ICL techniques similar to those described by Xu et al. (2023) and Zhang et al. (2023a). Specifically, we experimented with two prompting strategies: a single-turn setting, where the model receives a single prompt containing all instructions and examples at once, and a multi-turn setting, where the model has a view of multiple conversational turns, each providing incremental demonstrations illustrating how the LLM "had parsed" a

review into quadruples. Our preliminary experiments indicated that adopting the single-prompt ICL approach of Bai et al. (2024) for ASQP resulted in inferior performance. Averaged across all datasets, the multi-turn style prompt increased the F-score by approximately 21% for Qwen and 14% for Gemini. In contrast, GPT showed comparable performance between single- and multi-turn settings (a difference of approximately 0.4%). We maintained identical header and closing prompts across experiments to compare the two approaches, varying only the demonstration examples provided in the middle.

To construct the ICL examples, we aimed to maximise the coverage of category and polarity combinations present in the training set. Specifically, we stratified the selection process to ensure that, whenever possible, each unique combination of category and polarity was represented at least once among the selected examples. To achieve this, we first shuffled the training set. We then iteratively selected examples such that each example contributed a new category-polarity combination until all available combinations were covered or the desired number of examples was reached. If the number of requested examples was less than the number of unique categories, we prioritised selecting examples such that each represented a distinct category. In cases where the number of requested examples exceeded the number of unique category-polarity combinations, we filled the remaining slots with randomly selected examples from the training set. This procedure ensured a diverse and representative set of ICL examples.

3.2.1 Prompt Crafting

Firstly, following the recommendations of Wang and Luo (2023), we employed a role-playing instruction for the LLM, designating it as an NLP assistant expert in ABSA. This approach required the LLM to provide precise answers and strictly adhere to the given instructions.

Subsequently, we adopted the prompt style of Xu et al. (2023) and Zhang et al. (2023a). We indicated whether an element must be extracted from the text or could be left 'NULL'. We then listed the possible categories of the domain, specified the polarity range, and showcased the format of a quadruple. Based on the number of demonstrations, we supplied a history of conversations to be used as context for the LLMs. Figure 2 shows an example with one demonstration.

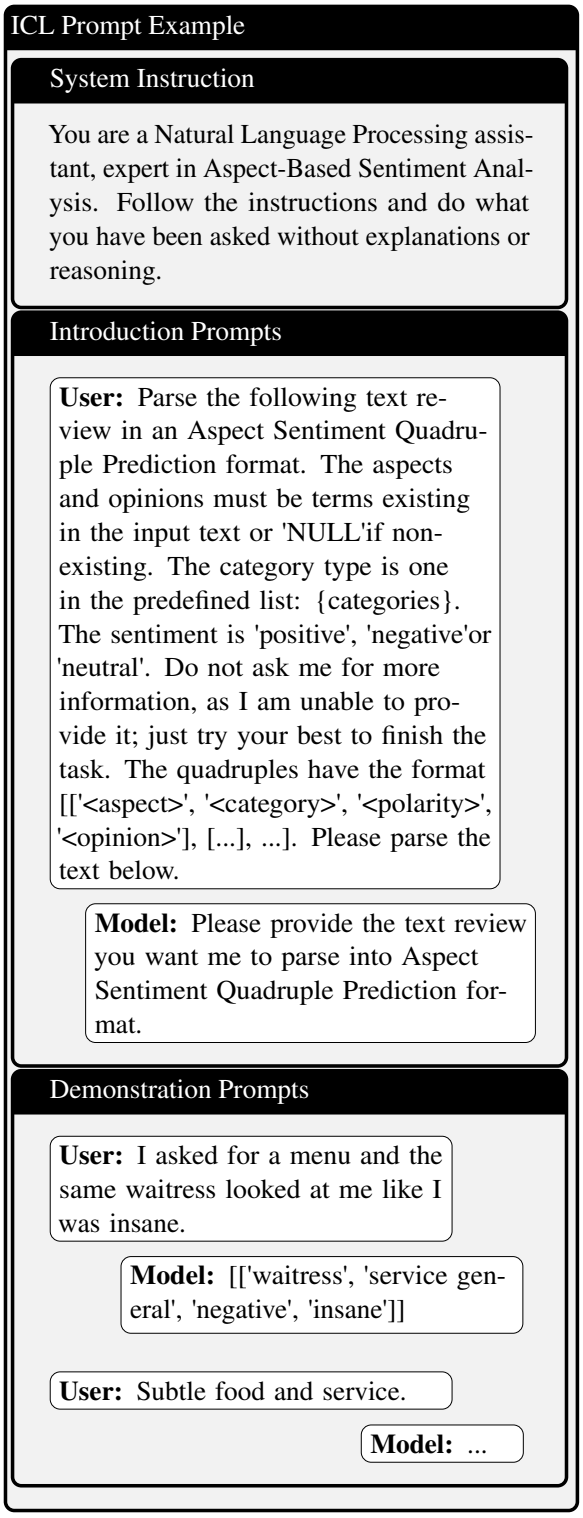

Figure 2: An ICL multi-turn prompt example. Firstly, the system instruction narrates the LLM how to act. Then the introduction prompt defines the task. Following, there is one demonstration, and then the model is expected to provide the quadruple for the last sentence.

3.3 CoT Agents

Our study aimed to determine the most effective sequence of thought processes for extracting sentiment information. Intuitively, the same pattern is seen in previous studies on related ABSA tasks (Fei et al., 2023; Wang and Luo, 2023), one might start by extracting aspects, followed by categories from a predefined list, then expressing opinions, and finally the polarity of those opinions. In our study, we considered all four permutations, resulting in 24 possible sequences for extracting these elements.

Our preliminary experiments suggested that extracting both the opinion and its polarity for a detected category in a review can be effectively accomplished in a single turn, rather than in two turns. Specifically, we found that directly prompting for the sentiment of a category within a review (with a prompt: "What is the sentiment of category <category> in the text <review>?") yields marginally better results when averaged across all datasets and models, compared to first extracting the opinion and subsequently asking for its sentiment. Consequently, we adopted the single-step approach to improve efficiency and reduce unnecessary computational costs, resulting us to 6 possible sequences.

3.3.1 Prompt Crafting

Our next step was to craft the necessary prompts to chain these permutations. Fei et al. (2023) and Wang and Luo (2023) showcase their prompt formats, both of which follow the same CoT order of elements: aspects, opinion, and polarity for the task of sentiment classification. However, their prompts are constructed differently. One approach uses prompts to ask the LLM each element one by one, whereas the other copies the generated answer from the LLM from the previous query and appends it to the next query. In our study, we aimed to balance these two methods in a new approach. Specifically, we sought to achieve the conciseness of the second method by including pointers to chain them together. For instance, if we first extracted the aspects and then wanted to extract the sentiments, our prompt would be: "List all word sequences that denote or link to a sentiment from the detected aspects. Sentiments:".

Additionally, similar to the ICL prompting format, we included a system instruction, which acted as a role-playing format and constrained the LLM to generate the most probable answers while scrutinising their verbosity. In the end, we added two demonstrations, similar to Figure 2, to demonstrate how a sentence ought to be parsed.

Number of:	Rest15	Rest16	Amazon	Hotels	Laptops	Shoes
categories	13	12	10	74	114	21
train sampl.	834	1264	1374	1438	2934	906
dev sampl.	209	316	194	203	326	116
test sampl.	537	544	395	414	816	125
quads	795	799	1657	3222	1161	518
impl. asp.	218	179	318	141	235	253
impl. op.	0	0	5	30	319	0

Table 1: Dataset statistics for the six datasets employed in our study. The 'quads' row denotes the number of resulting quadruples for the test set. The implicit aspects (impl. asp.) row refers to the number of quadruples in the test set where the first element, the aspect, is not explicitly mentioned in the text, resulting in a value of 'NULL', but still implying a category. Similarly, for the implicit opinions (impl. op.).

4 Experiments

4.1 Datasets

For our experiments, we included six English datasets. Among these, we selected the Rest15 and Rest16 restaurant review ASQP datasets, which were initially introduced by Pontiki et al. (2015, 2016) and subsequently curated by Peng et al. (2020), Wan et al. (2020), and Zhang et al. (2021a). These datasets are among the most extensively evaluated in the ASQP domain (see, for example, Wang et al., 2023; Zhang et al., 2024; Hu et al., 2023; Yu et al., 2023; Hu et al., 2022; Zhou et al., 2023, inter alia).

In addition, we included two whole-review ASQP datasets: the Amazon Fine Foods Reviews and the Hotel Reviews from TripAdvisor (Chebolu et al., 2024). We also incorporated the ACOS Laptops dataset (Cai et al., 2021), which consists of sentence-level annotations. ACOS, or aspect-category-opinion-sentiment quadruple extraction, produces the same output as ASQP but focuses on extracting implicit aspects and opinions. Finally, we included the Shoes-ACOSI dataset (Peper et al., 2024), which is annotated at the whole-review level. Shoes-ACOSI is a quintuple ACOS-type dataset that introduces an additional flag (I) to indicate whether an opinion is implicit or explicit; this flag is not considered in the present study. In Table 1 we provide the datasets' statistics.

4.2 LLM Models

For our current study, we aimed to evaluate popular LLMs that provide accessible APIs or are easily

deployable with rapid response times, thereby eliminating the need for extensive infrastructure on the user's end. Specifically, we focused on models offered by OpenAI, Google, and Alibaba, selecting their respective distilled flagship models. The three models used in our study are OpenAI's GPT-4o[2] (OpenAI et al., 2024), Google's Gemini Flash[3] (Gemini et al., 2024), and Alibaba's Qwen3-30B-A3B (Yang et al., 2025).

Their enhanced efficiency and reduced computational requirements motivated the decision to utilise distilled models. Model distillation refers to compressing a larger neural network into a more compact model, which retains much of the original model's effectiveness while operating with greater speed and lower resource consumption (Zhu et al., 2023). The use of distilled models enabled us to conduct our evaluations more rapidly and cost-effectively, thereby facilitating a broader and more practical assessment of the capabilities and limitations of these leading LLMs.

4.3 Evaluation

Aspect Sentiment Quad Prediction A quadruple is considered correctly predicted when all its elements exactly match the ground-truth quadruple. Despite being a stringent criterion, this method aligns with those employed in previous studies (Hu et al., 2022; Yu et al., 2023; Wang et al., 2023; Zhang et al., 2024; Zhou et al., 2023). Consequently, we report precision, recall, and F1-score based on micro-averaged aggregation.

Aspect	Category	Polarity	Opinion

Aspect Sentiment Triplet Extraction For comparison, we also evaluated the aspect sentiment triplet extraction (ASTE) performance by ignoring the category element from the quadruple. We reported the same metrics as mentioned above.

Aspect	Polarity	Opinion

Aspect Category Sentiment Analysis The task of aspect-category sentiment analysis (ACSA) focuses solely on extracting the category and polarity of a quadruple. This task is best characterised as a multi-label classification problem, as the categories are predefined per domain and the polarity values range across positive, neutral, and negative. We adhered to the same metrics as those used in the previously mentioned evaluations.

Category	Polarity

4.4 Results & Discussion

4.4.1 In-Context Learning

For the ICL evaluations, we tested with 0, 2, 10, 20, 50, 100, and all available samples from the training set as demonstrations. We adhered to our filtering method specifically to prioritise samples representative of both category and polarity.

In our comparative analysis of the three models, GPT-4o notably achieved moderate F-scores even with minimal to no demonstrations, demonstrating robust performance out of the box, as shown in Figure 3. In contrast, the Gemini model struggled to deliver satisfactory results under similar conditions. This underperformance can be attributed to issues in parsing input into quadruples and inconsistencies in the formation of these quadruples, which resulted in varying sizes and compromised the model's effectiveness.

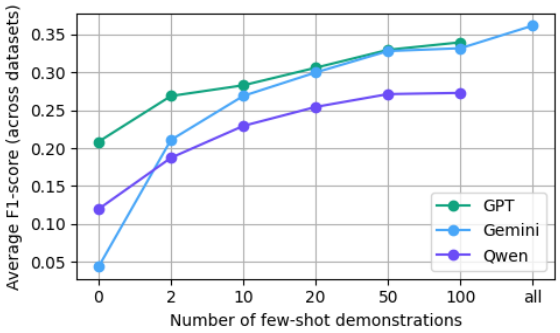

Figure 3: Average F-Scores for each model across the datasets.

All models exhibited a steady increase in performance as more demonstrations were provided, surpassing benchmarks set by earlier GPT versions (Xu et al., 2023; Zhang et al., 2023a). According to Figure 3, the most significant improvements in performance for all models were observed with increases in the number of demonstrations up to 20. This trend could indicate that providing 20 or more ICL demonstrations can yield satisfactory results for any given task.

Moreover, Gemini was capable of having the

[2]GPT-4o-2024-08-06
[3]Gemini-2.0-flash

whole training set as ICL demonstrations, whereas GPT and Qwen did not have the context capacity, which seemed resourceful as seen in Figure 3.

For reference, we include previously reported results from the literature(see Section 2), and compare them to the ICL agents of each model using 20, 100, and all training instances as demonstrations in Table 2. Overall, we achieve performance comparable to fine-tuned approaches, with increased samples leading to improved scores. The Laptops dataset contains more categories; therefore, it requires more examples, and we observe that using all available training samples proves particularly beneficial.

The ICL method faces the class imbalance issue since we input the raw instances directly. In contrast, various ASQP methodologies in the literature incorporate different data augmentation techniques to counter this imbalance. Adopting such techniques could improve ICL.

Performance across the three tasks revealed apparent differences in difficulty and responsiveness to ICL. On average, models achieved the highest performance on the ACSA task, followed by ASTE and ASQP. ASTE exhibited the most significant average improvement when increasing from zero-shot to 100-shot in-context learning, followed by ASQP. These results suggest that while models are relatively proficient at multi-label classification (ACSA), they face greater challenges in accurately identifying and extracting specific text segments, particularly in the more complex ASQP and ASTE tasks. This difficulty may be attributed to annotator biases, as extracted spans (i.e., aspects or opinions) must exactly match the ground truth annotations. Consequently, models may require more annotated examples to capture annotator-specific span preferences and thus achieve improved performance effectively.

4.4.2 CoT Agents

We tested three-element permutations (i.e., aspect, sentiment, category) for each model. Figure 4 presents the results of each permutation, averaged across datasets for each model for the ASQP task. As shown, GPT achieves the highest performance with an F-score of 0.170 (highlighted in bold), followed closely by Gemini with 0.145. GPT demonstrates greater robustness across different CoT permutations, possibly due to its larger parameter size.

The results indicate that, on average, starting with sentiment yields better performance, whereas

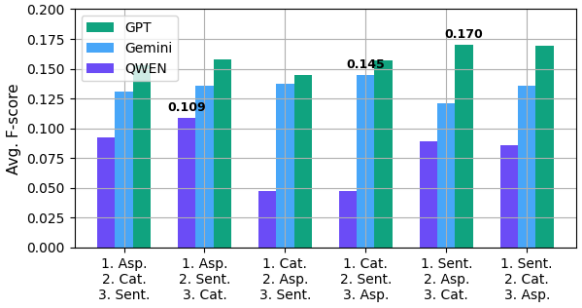

Figure 4: The performance of ASQP by different CoT agents across datasets and models. The x-axis labels show the sequence of CoT. For instance, extracting the sentiments first, then the categories, and finally the aspects would be represented as 1. Sent. 2. Cat. 3. Asp. The numbers in bold are the top averaged CoT F-Score for each model.

starting with aspects yields the lowest performance. This finding challenges the previously intuitive assumption that beginning with aspects would be beneficial. Further analysis is required to determine whether this improvement arises from dependency parsing facilitating extraction, or if it reflects an inherent bias in large language models toward responding more effectively to emotional and negative phrases (Wang et al., 2024), as well as to identify whether there are optimal CoT sequences for specific cases.

When comparing the ICL approach using two samples to the CoT approach, which also utilised two samples, we observed that the ICL approach still achieved better results than any CoT variant, indicating that further investigation in this direction is required. One possible explanation is that, for the ASQP task, which is heavily context-dependent, the ICL approach may be more effective at capturing annotation biases associated with extracting information from text. Our hypothesis regarding span annotation bias warrants further exploration, and alternative evaluation methods should be examined—particularly those employing less stringent quad matching criteria—to determine whether alternative ICL approaches based on zero-shot or few-shot settings can yield improved performance.

4.5 Hyper-Parameters & Generation Issues

For our experiments, we set the temperature hyper-parameter to zero and employed greedy decoding. We accessed the models through their APIs, and for Qwen specifically, we utilised vLLM (Kwon et al., 2023) running on an H100 SXM GPU.

In general, the LLMs produced consistent

Method	ASQP												ACOS					
	Rest15			Rest16			Amazon_FF			Hotels			Laptops			Shoes		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
GAS	45.3	46.7	46.0	54.5	57.6	56.0	–	–	23.6	–	–	44.6	–	–	–	–	–	–
Para	46.2	47.7	46.9	56.6	59.3	57.9	–	–	**26.1**	–	–	**46.0**	–	–	–	–	–	–
DLO	47.1	49.3	48.2	57.9	**61.8**	59.8	–	–	–	–	–	–	–	–	–	–	–	–
DAST	50.0	49.7	49.8	**62.8**	60.3	**61.5**	–	–	–	–	–	–	–	–	–	–	–	–
GenDA	49.7	50.3	50.0	60.1	61.7	60.9	–	–	–	–	–	–	–	–	–	–	–	–
MvP	–	–	51.0	–	–	60.4	–	–	–	–	–	–	44.6	43.6	44.1	**21.4**	18.2	**19.7**
GEN-S-N	–	–	–	–	–	–	–	–	24.0	–	–	43.8	**46.7**	**45.8**	**46.2**	19.6	**18.6**	19.1
Qwen 20	32.9	38.0	35.3	41.1	47.7	44.1	13.3	15.1	14.1	24.7	24.4	24.6	21.4	22.2	21.8	13.7	11.8	12.7
Qwen 100	38.6	43.9	41.1	41.3	46.7	43.8	15.5	17.7	16.5	26.4	24.7	25.5	22.4	22.3	22.4	15.4	13.5	14.4
GPT 20	38.9	43.7	41.1	49.8	54.7	52.1	19.9	18.7	19.3	35.1	30.9	32.9	22.3	23.7	23.0	16.4	14.1	15.2
GPT 100	47.6	49.7	48.6	54.8	57.5	56.1	22.3	20.0	21.1	38.3	32.7	35.3	27.7	27.0	27.4	16.8	13.9	15.2
Gem 20	43.7	45.3	44.5	52.4	54.7	53.5	17.7	15.6	16.6	35.4	20.9	26.3	25.3	24.8	25.0	16.3	12.2	13.9
Gem 100	50.8	50.6	50.7	53.2	55.7	54.4	22.4	20.1	21.1	36.7	17.8	24.0	32.0	29.2	30.5	20.7	16.2	18.2
Gem All	**53.2**	**52.3**	**52.8**	60.0	61.4	60.7	19.5	16.6	17.9	36.1	19.5	25.3	43.2	40.4	41.8	21.8	15.8	18.3

Table 2: Precision, recall, and F1-score metrics (%) on the ASQP and ACOS datasets. The overall best value for each column is shown in bold, while the best result among our ICL variants is underlined. All baseline systems are summarised in Section 2.1; their references are indicated in the Method column by the corresponding acronyms. For MvP, we report the model variant trained solely on the target dataset. Scores for Amazon and Hotels are taken from Chebolu et al. (2024), those for Laptops and Shoes from Peper et al. (2024), and the Rest15 and Rest16 results from the respective original papers.

quadruples according to the specified format, comprising four elements separated by commas and enclosed in square brackets. GPT demonstrated greater consistency, even in a zero-shot ICL setup, by following simple instructions on generation. In contrast, Gemini performed significantly better when provided with at least two ICL demonstrations. Its outputs generally exhibited slightly higher recall than precision, highlighting the model's tendency towards verbosity.

Furthermore, regarding the CoT process, there were instances where the LLM struggled to generate an answer due to the brevity of the text. Examples of such text include phrases like "awesome," "try it!," "No comparison," or brief descriptions of situations implying sentiment. Nonetheless, these instances were fewer than ten in each case.

5 Conclusion

In this study, we extensively investigated ICL applied to state-of-the-art distilled LLMs. Our analysis demonstrated that employing multiple ICL demonstrations enables these models to replicate previously reported benchmarks achieved by encoder-decoder transformer architectures and significantly outperform earlier model generations. Gemini's extended context proved beneficial by allowing the inclusion of the entire training dataset, subsequently enhancing F-score metrics—particularly valuable for datasets containing multiple categories. Additionally, the chat-based (multi-turn) ICL approach emerged as a superior method for presenting few-shot demonstrations. Notably, GPT exhibited an exceptional capability to format outputs directly using a zero-shot approach. All models showed incremental improvements in F-score correlated with increased sample sizes but would start showing signs of plateau after 20 examples.

Furthermore, we conducted experiments with CoT prompting by evaluating performance across different CoT sequences. Contrary to prevailing assumptions that extracting the aspect first is optimal, we found that initiating CoT with alternative elements can yield improved scores. However, CoT generally did not achieve high overall results, indicating that ICL may be a more effective technique. We hypothesise that a major hindrance to overall ASQP performance could be the inherent annotation bias involved in extracting text spans for quadruples, a factor that may favour fine-tuned approaches over few-shot and CoT-based methods.

References

Yinhao Bai, Zhixin Han, Yuhua Zhao, Hang Gao, Zhuowei Zhang, Xunzhi Wang, and Mengting Hu. 2024. Is compound aspect-based sentiment analysis addressed by LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7836–7861, Miami, Florida, USA. Association for Computational Linguistics.

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, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess,

Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. Roast: Review-level opinion aspect sentiment target joint detection for absa.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *ArXiv*, abs/2301.00234.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

Team Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, and et al. (1037 additional authors not shown). 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13481–13494, Toronto, Canada. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott

Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, and Shawn Jain et al. (181 additional authors not shown). 2024. GPT-4 technical report.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8600–8607.

Joseph Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6089–6095, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joseph J Peper, Wenzhao Qiu, Ryan Bruggeman, Yi Han, Estefania Ciliotta Chehade, and Lu Wang. 2024. Shoes-ACOSI: A dataset for aspect-based sentiment analysis with implicit opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15477–15490, Miami, Florida, USA. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z. Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9122–9129.

An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023. Generative data augmentation for aspect sentiment quad prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 128–140, Toronto, Canada. Association for Computational Linguistics.

Xu Wang, Cheng Li, Yi Chang, Jindong Wang, and Yuan Wu. 2024. Negativeprompt: Leveraging psychology for large language models enhancement via negative emotional stimuli.

Yajing Wang and Zongwei Luo. 2023. Enhance multi-domain sentiment analysis of review texts through prompting strategies. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 1–7.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xiancai Xu, Jia-Dong Zhang, Rongchang Xiao, and Lei Xiong. 2023. The limits of chatgpt in extracting aspect-category-opinion-sentiment quadruples: A comparative analysis. *ArXiv*, abs/2310.06502.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Yongxin Yu, Minyi Zhao, and Shuigeng Zhou. 2023. Boosting aspect sentiment quad prediction by data augmentation and self-training. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing-Quan Liu, Sinno Jialin Pan, and Lidong Bing. 2023a.

Sentiment analysis in the era of large language models: A reality check. *ArXiv*, abs/2305.15005.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. 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 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

Wenyuan Zhang, Xinghua Zhang, Shiyao Cui, Kun Huang, Xuebin Wang, and Tingwen Liu. 2024. Adaptive data augmentation for aspect sentiment quad prediction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11176–11180.

Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou, and JunBo Yang. 2023. A unified one-step solution for aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12249–12265, Toronto, Canada. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *ArXiv*, abs/2308.07633.