Prompted Aspect Key Point Analysis for Quantitative Review Summarization

An Quang Tang[*♣] **Xiuzhen Zhang**[*♣] **Minh Ngoc Dinh** [♣] and **Erik Cambria**[◇]
[♣]RMIT University, Australia
[◇]Nanyang Technological University, Singapore
s3695273@rmit.edu.vn, xiuzhen.zhang@rmit.edu.au
minh.dinh4@rmit.edu.vn, cambria@ntu.edu.sg

Abstract

Key Point Analysis (KPA) aims for quantitative summarization that provides key points (KPs) as succinct textual summaries and quantities measuring their prevalence. KPA studies for arguments and reviews have been reported in the literature. A majority of KPA studies for reviews adopt supervised learning to extract short sentences as KPs before matching KPs to review comments for quantification of KP prevalence. Recent abstractive approaches still generate KPs based on sentences, often leading to KPs with overlapping and hallucinated opinions, and inaccurate quantification. In this paper, we propose Prompted Aspect Key Point Analysis (PAKPA) for quantitative review summarization. PAKPA employs aspect sentiment analysis and prompted in-context learning with Large Language Models (LLMs) to generate and quantify KPs grounded in aspects for business entities, which achieves faithful KPs with accurate quantification, and removes the need for large amounts of annotated data for supervised training. Experiments on the popular review dataset Yelp and the aspect-oriented review summarization dataset SPACE show that our framework achieves state-of-the-art performance. Source code and data are available at: https://github.com/antangrocket1312/PAKPA

1 Introduction

With the sheer volume of reviews, it is impossible for humans to read all reviews. Although the star ratings aggregated from customer reviews are widely used by E-commerce platforms as indicators of quality of service for business entities (McGlohon et al., 2010; Tay et al., 2020), they can not explain specific details for informed decision-making. Early review text summarization studies focused only on capturing important points with high consensus (Dash et al., 2019; Shandilya et al., 2018), yet overlooked minor ones and also were unable to measure the opinion prevalence.

Key Point Analysis (KPA) is proposed to summarize opinions in review comments into concise textual summaries called Key Points (KPs), and quantify the prevalence of KPs. KPA studies were initially developed for argument summarization (Bar-Haim et al., 2020a), and then adapted to business reviews (Bar-Haim et al., 2020b, 2021). Most KPA studies adopt the extractive approach, which employs supervised learning to identify informative short sentences as key points (KPs), often leading to non-readable and incoherent KPs. Recently, KPA studies apply abstractive summarization to paraphrase and generate KPs from comments (sentences) (Kapadnis et al., 2021; Li et al., 2023). Still, KPs are generated based on sentences and often contain unfaithful and overlapping opinions, and inaccurate quantity for their prevalence.

In this paper, we propose Prompted Aspect Key Point Analysis (PAKPA). Different from previous sentence-based KPA studies, our system employs aspect-based sentiment analysis (ABSA) to identify aspects in comments as the opinion target and then generate and quantify KPs grounded in aspects and their sentiment. Importantly, to address the issue of scarce annotated data for supervised training in existing studies, we employ prompted in-context learning with LLMs for ABSA extraction and KP generation. By integrating ABSA into the prompted summarization process, we aim to mitigate hallucination by guiding LLMs to produce KPs aligned with the common aspects shared by reviews. Table 1 shows the top positive KPs generated by PAKPA, ranked by their prevalence, for reviews of a hotel business entity.

Our contributions are two-fold. To our best knowledge, we are the first to employ prompted in-context learning for abstractive KPA summarization of reviews, which removes supervised training using large amount of annotated data. Secondly,

[*]First two authors equally contributed to this work.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10691–10708

Key Point	Prevalence	Matching Comments
Friendly and helpful staff.	46	From the minute I walked in the door the staff has treated me with such kindness and respect, there are not enough words I can say of how much gratitude I have for the staff here.
		It's a fine hotel, pretty basic, and all of the staff members we encountered were quite friendly.
Excellent location for business travelers.	37	Its a great location for business travelers since I can stay here and always be going the opposite way of traffic.
		*This hotel has an IDEAL location, the shuttle was perfect, and we got a great deal on priceline.
Clean and comfortable rooms.	32	*You will enjoy the breathtaking views from your spacious, clean and comfortable room.
		The rooms were nice and clean and the bed was comfortable.
Convenient and helpful shuttle service.	12	*This hotel has an IDEAL location, the shuttle was perfect, and we got a great deal on priceline.
		I would also like to give a shout out to the terrific shuttle drivers Jeff and Rod who were so great to my family group.
Amazing view of Vanderbilt football stadium or The Parthenon.	10	*You will enjoy the breathtaking views from your spacious, clean and comfortable room.
		Then the surprise came when we opened the curtains to see a full on view of the Vanderbilt football stadium.

Table 1: Top positive key points of a hotel business entity generated by PAKPA. For each key point, we show the prevalence, i.e., number of matching comments, and two randomly selected matching comments.

our approach of integrating aspect-based sentiment analysis (ABSA) into KPA for fine-grained opinion analysis of review comments ensures generating KPs grounded in aspects for business entities and more accurate matching of comments to KPs, resulting in faithful KPs for distinct aspects as well as more accurate quantification of KP prevalence.

2 Related Work

Based on the form of summaries, review summarization studies can be broadly grouped into three classes: key point analysis, aspect-based structured summarization, and text summarization. Additionally, we reviewed the forefront of prompted in-context learning for review (text) summarization.

2.1 Key Point Analysis

Developed initially to summarize arguments (Bar-Haim et al., 2020a), KPA was later adapted to summarize and quantify the prevalence of opinions in business reviews (Bar-Haim et al., 2020b, 2021; Tang et al., 2024). The majority of KPA studies focus on extracting short sentences as salient KPs from arguments or review comments, and then matching KPs to comments to quantify their prevalence. These works employ supervised learning to train models to identify informative KPs, which require large volumes of annotated training data, and the resulting KPs may not be succinct textual summaries and may not represent distinct salient opinions either. An exception is ABKPA (Tang et al., 2024), which adopts an aspect-based approach to produces concise KP summaries. Still, the approach produced non-informative KPs due to its extractive mechanism, and requires supervised learning to train models for matching KPs to comments for KP quantification.

Recently, abstractive KPA studies have proposed generating KPs using abstractive text summarization approaches for arguments rather than reviews. Kapadnis et al. (2021) initially proposes to generate KPs for each argument (sentence) before selecting representative ones based on ROUGE scores. However, the technique basically rephrases arguments as KPs. Li et al. (2023) then suggests clustering similar arguments, based on their *contextualised embeddings*, before using an abstractive summarization model to generate concise KP condensing salient points. But the approach is not feasible for reviews because review comments can contain multiple opinions on different aspects of business entities, and clustering comments by only their sentence-level embeddings cannot accurately identify distinct KPs on different aspects, leading to inaccurate quantification. Moreover, KPA for reviews remains an open challenge due to the lack of large-scale annotated dataset for KP generation.

2.2 Aspect-based Structured Summarization

Early works from the data mining community focus on aspect-based structured summarization, which applies Aspect-Based Sentiment Analysis (ABSA) to extract, aggregate, and organize review sentences into a hierarchy based on features (i.e. aspects) such as food, price, service, and their sentiment (Hu and Liu, 2004; Ding et al., 2008; Popescu and Etzioni, 2007; Blair-Goldensohn et al., 2008; Titov and McDonald, 2008). These works lack textual explanation and justification for the aspects and their sentiment.

2.3 Text Summarization

More broadly, document summarization is an essential topic in the Natural Language Processing community, aiming to produce concise textual summaries capturing the salient information in source documents. While extractive review summarization approaches use surface features to rank and extract salient sentences into summaries (Mihalcea and Tarau, 2004; Angelidis and Lapata, 2018; Zhao and Chaturvedi, 2020), abstractive techniques use sequence-to-sequence models (Chu and Liu, 2019; Suhara et al., 2020; Bražinskas et al., 2020b,a; Zhang et al., 2020) to paraphrase and generate novel words not in the source text. Still, none of these studies can capture and quantify the diverse opinions in reviews.

2.4 Prompted Opinion Summarization

For generation of textual summaries, recent studies successfully applied summarization prompt on LLMs to generate review summaries (Bhaskar et al., 2023; Adams et al., 2023). Notably, to overcome the length limit for the input text from `GPT3.5`, Bhaskar et al. (2023) splits the input into chunks and summarize them recursively to achieve the final textual summary. Nevertheless, these studies still leave unexplored the use of in-context learning in LLMs for quantitative summarization, particularly in presenting and quantifying the diverse opinions in reviews.

3 Methodology

Figure 1 illustrates our PAKPA framework with examples. Given reviews for a business entity, PAKPA performs KPA for reviews and generates KPs of distinctive aspects and quantities measuring the prevalence of KPs. PAKPA consists of three components:

- *Prompted Aspect-Based Sentiment Analysis (ABSA) of comments:* extracts the aspect terms and sentiment – positive or negative – for each review comment (sentence),
- *Aspect sentiment-based comment clustering:* clusters comments sharing similar aspects and sentiments
- *Prompted aspect KP generation:* generates aspect KPs from comment clusters.

Core to our framework is to employ ABSA of review comments to identify aspect terms in reviews and predict their sentiment, which sets the basis for clustering comments based on aspects and for further generation of aspect-oriented KPs. This idea is inspired by the early aspect-based structured summarization studies (Hu and Liu, 2004; Ding et al., 2008), which aggregates review comments by their sentiment toward common aspects for more accurate quantification of opinions. Importantly, prompted in-context learning strategies are employed for aspect-based sentiment analysis of review comments, and aspect-oriented KP generation and quantification.

3.1 Prompted Aspect-based Sentiment Analysis of Comments

We designed and employed prompted in-context learning for LLMs to extract ABSA from reviews. The task is to predict (a, s) pairs – (a)spect term, and (s)entiment (positive, neutral or negative) – for each review sentence. We developed a simple prompting strategy based on OpenAI [1]'s prompt engineering guidelines. Our prompts are structured into five parts, as shown in Table 2: 1) Context of the review comment to be analyzed; 2) Definition of the ABSA task and the expected elements to retrieve; 3) Request for the LLM to provide the label in a JSON format; 4) Few-shot (18) examples to guide the LLM to generate the desired type of response; and 5) Review comment for ABSA predictions. Experiments (Section 4.3) show that our prompted approach achieved reasonable performance on the aspect extraction and sentiment prediction tasks compared to supervised ABSA models.

3.2 Aspect Sentiment-based Comment Clustering

Clustering comments directly based on their identical aspect terms can be highly overlapping because there are semantically similar aspect terms among the clusters. We aim to construct clusters

[1] `https://platform.openai.com/docs/guides/prompt-engineering`

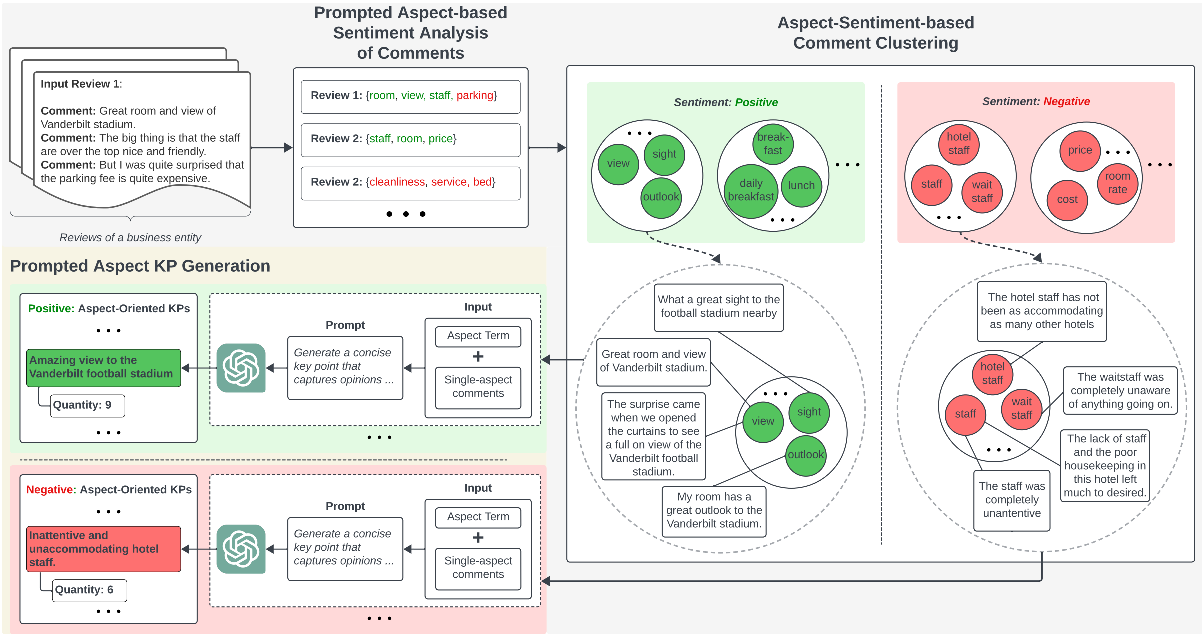


Figure 1: The PAKPA framework

Prompt for ABSA of Comments	Prompt for Aspect Key Point Generation
You will be provided with a review sentence delimited by triple quotes. A review sentence usually covers the customer opinions expressed on different aspects of a product or service. You are tasked to perform Aspect–based Sentiment Analysis to extract the user sentiments expressed on different aspects in the review. Formally, we define subtask of extracting the aspects it corresponding sentiments as Aspect Extraction and Aspect Sentiment Classification: – Aspect Extraction: Identifying aspect targets in opinionated text, i.e., in detecting the specific aspects of a product or service the opinion holder is either praising or complaining about. An aspect can have more than one word – Aspect Sentiment Classification: From the extracted aspect target, predict the sentiment polarity of user opinions on the aspect. The sentiment polarity value can be: " positive", "neutral", and "negative". Provide the answer in JSON format with the following keys: aspect, sentiment	You will be provided with a list of user review comments delimited by triple quotes, and a list of common aspects shared by those reviews delimited by triple quotes The comments in the list has been clustered by some common aspects and sentiment. You are guided to generate a concise key point that captures opinions on the most popular aspect across the input comments, and also accomodate the provided list of common aspect. Note that the generated key points must describe the opinion in only ONE aspect only and must not discuss multiple aspects. The generated key points must have 3–5 tokens. Perform the following actions to solve this task: – Identify the single and general aspect (e.g. atmosphere) that are common across the input aspects terms – On the identified aspect, find the salient points of opinions mentioning that aspect across the input comments Some invalid examples of key points with multiple aspects that must be avoided: – "Enjoyable atmosphere with great music and live entertainment.", rather it should be " The atmosphere is very enjoyable." – "Excellent wine selection and enjoyable atmosphere.", rather it should be "The wine selection is great."

Table 2: Prompts for "ABSA of Comments" and "Aspect Key Point Generation" of the PAKPA framework. Full prompts with few-shot examples are provided in Appendix A

of comments such that comments of the same cluster will share the same aspect and sentiment, and each cluster has distinct aspect and sentiment from the other. To achieve this object, we leverage the (aspect, sentiment) pairs extracted from comments by our Prompted ABSA process (Section 3.1). We propose a greedy algorithm to construct clusters for comments, based on their sentiment and semantically similar aspect terms.

Let $R_e = \{r_i\}_{i=1}^{|R_e|}$ denotes a set of review comments on a business entity e. First we start by applying prompted ABSA (discussed in Section 3.1) on r to extract possible (a)spect terms and the (s)entiment in a comment as a list of $(a,\ s)$ pairs. Formally, this can be defined as $O_r = \{(a_m, s_m)\}_{m=1}^{|O_r|}$, where s_m is the sentiment polarity of the m-th aspect in r. (*positive*, *neutral*, or *negative*). Note that hereafter we filter all neutral sentiments in O_r. We then aggregate all aspect terms (a_m) of the same sentiment in $r_i \in R_e$ into A_{pol}, where pol is either the positive or negative.

Given a A_{pol} of R_e, we first rank all aspects by descending order of their frequency in R_e. Then we start with an empty $\mathbf{C}$, and iterate through every aspect in A_{pol}. For every aspect, we further iterate through every existing cluster in $\mathbf{C}$ and calculate the average cosine similarity score to all included aspects of the cluster. Finally, only the aspect with the highest average cosine similarity score is added to the cluster with a threshold (λ) above 0.55, otherwise, a new cluster is created. As shown in Figure 1, an example of semantically similar aspect terms is *view*, *sight*, and *outlook*, which can be grouped into a cluster.

We employ SpaCy (Honnibal et al., 2020) to calculate the cosine similarity between aspect terms to form clusters. Finally, comments sharing similar aspects, now grouped into clusters, are aggregated to become the input for the upcoming KP Generation stage, and the size of clusters is the quantity measuring the prevalence of KPs.

3.3 Prompted Aspect-oriented KP Generation

Unlike existing studies that rely on supervised text generation (Li et al., 2023), we achieve Key Point Generation (KPG) by prompting an LLM to generate concise, distinct KPs from clusters of comments with the semantically similar aspect terms. Our main idea is that semantically similar aspect terms of a cluster of comments can be a good signal to infer a high-level and more general aspect-oriented textual description as the KP. Specifically, we designed the prompt for Aspect KPG based on simple prompting strategies suggested by the OpenAI prompt engineering guideline [2] to write clear instructions to prompt the model. Our prompt is structured into six parts, as shown in Table 2: 1) Context of the KPG input to be summarized; 2) Definition of the Aspect KPG task and the output requirement; 3) Summarization steps to guide the LLM to infer the general aspects from the cluster's aspect terms and then generate aspect-oriented KP; 4) One-shot example to guide the LLM to generate the desired type of response; 5) Guiding the LLM through invalid generation examples to avoid, along with preferred correction for practicing; and 6) KPG input for summarization. We provide details of the prompt on LLMs for aspect-based KPG in Listing 2 (Appendix A).

4 Experiments

4.1 Implementation Details and Baselines

PAKPA was implemented with different LLMs, including open-source models `Vicuna-7B`, `Mistral-7B`, and the commercial model `GPT3.5`. Note that we employed open-source LLMs throughout PAKPA for both the ABSA stage and the KP generation stage, but we did not apply `GPT3.5` for the ABSA stage due to the exorbitant cost for thousands of reviews as the input context. We benchmark PAKPA against various baselines for extractive KPA, abstractive KPA, and the recent prompted opinion summarization.

Extractive KPA: We compare PAKPA against two latest extractive KPA systems **RKPA-Base** (Bar-Haim et al., 2021) and **ABKPA** (Tang et al., 2024). RKPA-Base is the first extractive KPA system for review summarization. It leverages a quality ranking model Gretz et al. (2020) to select KP candidates, and integrates sentiment analysis and collective key point mining into matching comments to the extracted KPs. ABKPA integrates ABSA into extracting and matching of KPs to comments for more precise matching and quantification of key points. We implement all models based on their default settings.

Abstractive KPA: We implemented two latest abstractive KPA systems **Enigma+** (Kapadnis et al., 2021) and **$\mathbf{SKPM_{Base}}$(IC)+** (Li et al., 2023). Enigma+ is adapted from the original Enigma framework to review data, which uses a Pegasus (Zhang et al., 2020) summarization model to generate KPs from comments, and selects the top 40 summaries based on their ROUGE scores. Similarly, $\mathrm{SKPM_{Base}(IC)+}$ is adapted for reviews, [3] employing `BERTopic` (Grootendorst, 2022) to cluster sentences and `Flan-T5` (Chung et al., 2022) to generate KPs. To fully adapt these works from arguments to reviews, we replace the topic and stance attribute in the input with business category and sentiment. We fine-tune all models using an annotated KP Matching dataset (Tang et al., 2024).

All above baselines were implemented either using the PyTorch module or the Huggingface transformers framework, and were trained on an NVIDIA GeForce RTX 3080Ti GPU.

Prompted Opinion Summarization: To evaluate the utility of KPA systems for textual summaries, we also compare them against the latest prompted opinion summarization model **Recursive GPT3-Chunking (CG)** (Bhaskar et al., 2023), which recursively chunks and prompts `GPT3.5` to generate textual summaries from user reviews. The final summary from this baseline is a paragraph rather than a list of KPs. For fair comparison, we follow the strategy of Bhaskar et al. (2023) by again

[2] `https://platform.openai.com/docs/guides/prompt-engineering`

[3] We reproduced this model based on the best configuration provided.

prompting GPT3.5 to split and rephrase the summary sentences into KPs. [4]

4.2 Datasets and Evaluation Dimensions

Datasets To evaluate both the textual quality and prevalence precision for KPs, we consider two popular datasets on business reviews, namely SPACE and YELP. **(1) SPACE**, featuring TripAdvisor hotel reviews, stands out as the only dataset providing human-annotated aspect-specific summaries and, therefore, serves as an ideal ground truth for evaluating our aspect-based generation of KPs in PAKPA. The dataset facilitates the evaluation of KP quality in capturing the main viewpoints of users across various aspects (e.g., location and cleanliness). **(2) YELP** is a widely used dataset for review summarization including a wider variety of business categories. This dataset is used to evaluate both the textual quality and quantification performance of KPs. Details of the datasets can be found in Appendix B.

Evaluation of KP Textual Quality with Aspect-Specific Ground Truth SPACE provides the reference summaries for this evaluation. Positive and negative summaries are evaluated separately. [5] We first perform a lexical comparison between generated KPs and the ground truth by computing the highest ROUGE score between generated and reference key points for each business entity and then average the maxima. Neverthless, KPs generated from abstractive KPA systems should not only be evaluated based on lexical similarity against ground truth summaries. We, therefore, employ the *set-level KPG evaluation* (Li et al., 2023), which explicitly measures the quality between two sets of generated and reference KPs based on their semantic similarity. For all business entities, we calculate the semantic similarity scores between the corresponding group of prediction and reference before macro-averaging their values to obtain *Soft-Precision (sP)* and *Soft-Recall (sR)*. While sP finds the reference KP with the highest similarity score for each generated KP, sR is vice-versa. We further define *Soft-F1* ($sF1$) as the harmonic mean between sP and sR as below, where f computes similarities between two individual key points, $\mathcal{A}$, $\mathcal{B}$ is the set of candidates and references and $n = |\mathcal{A}|$ and $m = |\mathcal{B}|$, respectively.

$$sP = \frac{1}{n} \times \sum_{\alpha_i \in \mathcal{A}} \max_{\beta_j \in \mathcal{B}} f(\alpha_i, \beta_j) \quad (1)$$

$$sR = \frac{1}{m} \times \sum_{\beta_i \in \mathcal{B}} \max_{\alpha_j \in \mathcal{A}} f(\alpha_i, \beta_j) \quad (2)$$

We use state-of-the-art semantic similarity evaluation methods BLEURT (Sellam et al., 2020) and BARTScore (Yuan et al., 2021) as f_{max}. For fair comparison, we select only KPs of at least 15 matched comments [6].

Evaluation of KP Faithfulness and Information Quality We manually evaluated the information quality of generated KPs considering 7 different dimensions, divided into two groups. The first group, inspired by previous KPA works (Friedman et al., 2021; Li et al., 2023), evaluates how well the generated KPs summarize the salient information from the corpus. It assesses KPs based on criteria REDUNDANCY, COVERAGE, and FAITHFULNESS (contrary to hallucination). The second group measures the utility of generated KPs for summarization, under four dimensions (Bar-Haim et al., 2021): VALIDITY, SENTIMENT, INFORMATIVENESS and SINGLE ASPECT. Details of these dimensions are in Appendix C.

We conducted pairwise comparison of KPs from different systems using Amazon Mechanical Turk (MTurk). Given a dimension for evaluation, each comparison involved choosing the better one from two sets of KPs, each taken from a different system. We selected the top 5 KPs by prevalence for each sentiment. Using the Bradley-Terry model Friedman et al. (2021), we calculated rankings from these comparisons among the models. We ensured high-quality annotations by employing workers with an approval rate of 80% or higher and at least 10 approved tasks, while hiding ABSA details and framework identities to prevent bias. For an example of an annotation, see Appendix D. We only performed this evaluation on the YELP dataset, as it contains reviews for five business categories, including hotel reviews of SPACE. Note also that to maintain a reasonable annotation cost, for every category in YELP, we select only one top popular business entity with the highest average number of KPs being generated across the models.

[4] Also known as the atomic value judgement (Bhaskar et al., 2023).

[5] we use SpaCy to perform sentiment analysis on every referenced summary sentence.

[6] approximately equivalent to the top 7-10 KPs with the highest prevalence across the models for each business.

Evaluation of KP Quantification Precision In this experiment, we evaluate the precision of different systems for matching KPs to comments to measure the prevalence of KPs, namely the KP quantification precision (Bar-Haim et al., 2021). Following previous studies (Bar-Haim et al., 2021; Tang et al., 2024), this was conducted on YELP for various business categories. Adjustments were made to some KPA baselines (e.g., RKPA-Base, ABKPA, Engima+) to ensure comparable *review coverage* (Bar-Haim et al., 2021) [7] by setting an appropriate threshold (t_{match}) for selecting the best-matching comment-KP pairs. For annotation, we employed 6 MTurk crowd workers per comment-KP pair, selecting only those with an 80% or higher approval rate and at least 10 approved tasks. Following Bar-Haim et al.'s, we exclude annotators with Annotator-$\kappa < 0$ for quality control. This score averages all pairwise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 5 other annotators. For labelling correct matches, at least 60% of the annotators had to agree that the match is correct, otherwise, it is incorrect.

Task	AE	ASC
Prompted `Vicuna7B`	80.5	77.14
Prompted `Mistral7B`	78.56	76.88
Snippext (Full training)	79.65	80.45
Snippext (Low-resource)	77.18	77.4

Table 3: The F1 score of prompted LLMs and the SOTA Snippext model (Miao et al., 2020) for ABSA, evaluated on the Aspect Extraction (AE) and Aspect Sentiment Classification (ASC) tasks.

4.3 Results

Evaluation of ABSA To evaluate the effectiveness of prompted LLMs compared to supervised approaches for ABSA, we benchmark their performance on two tasks, namely Aspect Extraction (AE) and Aspect Sentiment Classification (ASC), from SemEval 2016 Task 5 (Pontiki et al., 2016) and SemEval 2014 Task 4 (Pontiki et al., 2014) respectively. Experimental results from Table 3 show that prompted LLMs achieved reasonable performance on AE and ASC tasks compared to the state-of-the-art (SOTA) ABSA model Snippext (Miao et al., 2020).

[7] Fraction of comments captured and quantified in the summary

Evaluation of KP Quality Table 4 presents our evaluation of the textual quality of KPs generated by different systems, focusing on their lexical and semantic similarity to the SPACE ground truth. In its best setting, our framework, PAKPA, outperforms other baselines across all metrics, capturing approximately 66% (sR = 0.66) of the viewpoints expressed in manually annotated aspect-specific summaries. Notably, SKPMBase(IC)+, despite its superiority over Enigma+ in argument summarization (Li et al., 2023), underperforms in generating quality KPs from reviews, as indicated by most metrics. This inferiority is attributed to SKPM-Base(IC)+'s vulnerability to hallucination when summarizing from a large set of comments, due to its reliance on limited supervised training data. Conversely, Enigma+, which generates KPs by rephrasing a single review sentence, maintains acceptable quality in its abstractive KP generation.

Our manual evaluation [8] on KP information quality further supports above findings. Table 5 highlights the Bradley Terry scores, measured by 7 information quality dimensions, of the KPs produced on YELP. Overall, on all 7 dimensions, PAKPA exhibits the highest and most stable performance. For summarizing the salient points, our framework outperforms other baselines significantly on COVERAGE (CV) and REDUNDANCY (RD), as it suggests that our approach captures more diverse opinions and also more effectively reduces redundancy in the KPs thanks to its aspect-based clustering and generation process. Importantly, PAKPA outperforms all baselines in FAITHFULNESS, more than doubling the effectiveness in reducing hallucinations compared to other abstractive summarization systems. For generating good KPs for reviews, PAKPA outperforms other baselines greatly on VALIDITY (VL), mainly because our approach uses LLMs to generate KPs that are aligned better with the expected format. Nevertheless, high scores SN, IN and SA also also shows that PAKPA can generate KPs with richful opinion information, expressing clearer sentiment and on more specific aspect than other baselines.

Performance of PAKPA Table 4 additionally presents the performance of PAKPA using different base LLMs on the SPACE dataset. Overall, implementing a combination of LLMs as base

[8] To maintain reasonable annotation cost, we only conducted manual evaluation on the best LLM configuration for PAKPA ($\text{PAKPA}_{\text{Vicuna7B+GPT3.5}}$), selected from Table 4.

	ROUGE			BARTScore			BLEURT		
	R-1	R-2	R-L	sP	sR	sF1	sP	sR	sF1
$\text{PAKPA}_{\text{Vicuna7B+GPT3.5}}$	**0.648**	**0.364**	**0.510**	0.74	0.66	0.70	**0.61**	**0.51**	**0.56**
$\text{PAKPA}_{\text{Mistral7B+GPT3.5}}$	0.588	0.341	0.453	0.73	0.64	0.68	0.58	0.50	0.54
$\text{PAKPA}_{\text{Mistral7B+Mistral7B}}$	0.531	0.269	0.440	0.74	**0.68**	**0.71**	0.58	0.50	0.54
$\text{PAKPA}_{\text{Vicuna7B+Vicuna7B}}$	0.515	0.231	0.371	0.73	0.66	0.70	0.61	0.49	0.54
Enigma+ (Kapadnis et al., 2021)	0.628	0.346	0.492	0.74	0.65	0.69	0.56	0.49	0.52
CG (Bhaskar et al., 2023)	0.416	0.205	0.406	0.73	0.56	0.63	0.52	0.45	0.48
$\text{SKPM}_{\text{Base}}$(IC)+ (Li et al., 2023)	0.335	0.139	0.318	0.67	0.58	0.62	0.38	0.36	0.37
RKPA-Base (Bar-Haim et al., 2021)	0.552	0.292	0.488	**0.75**	0.59	0.66	0.59	0.46	0.52
ABKPA (Tang et al., 2024)	0.442	0.245	0.422	0.74	0.63	0.68	0.56	0.46	0.51

Table 4: (SPACE) Textual quality evaluation of generated KPs with aspect-specific ground truth. sP, sR and sF1 refer to Soft-Precision, Soft-Recall, and Soft-F1 respectively based on set-level evaluation method. Statistical analysis on all metrics shows that PAKPA significantly outperforms the baselines (paired t-test $p << 0.05$).

	CV	FF	RD	VL	SN	IN	SA
$\text{PAKPA}_{\text{Vicuna7B+GPT3.5}}$	**28.44**	**26.56**	**25.34**	**35.23**	**31.11**	**25.9**	**24.8**
Enigma+ (Kapadnis et al., 2021)	11.06	11.17	14.7	9.99	9.54	13.49	17.52
CG (Bhaskar et al., 2023)	15.12	12.84	15.73	10.36	14.6	12.59	10.79
$\text{SKPM}_{\text{Base}}$(IC)+ (Li et al., 2023)	9.94	12.41	13.28	7.7	8.87	13.04	9.34
RKPA-Base (Bar-Haim et al., 2021)	16.20	22.28	15.73	22.91	20.75	21.02	18.77
ABKPA (Tang et al., 2024)	19.24	14.74	15.21	13.81	15.12	13.96	18.77

Table 5: (YELP) Information quality evaluation of generated KPs by different dimensions. Reported are the Bradley Terry scores of 7 dimensions, from left to right, COVERAGE, FAITHFULNESS and REDUNDANCY, VALIDITY, SENTIMENT, INFORMATIVENESS, SINGLEASPECT. A visual overview can also be found in Figure 2 (Appendix F)

	Arts	**Auto**	**Beauty**	**Hotels**	**Rest**	**Avg.**
$\text{PAKPA}_{\text{Vicuna7B+GPT3.5}}$	**0.98**	**0.93**	**0.96**	**0.94**	**0.94**	**0.95**
ABKPA	0.80	0.86	0.80	0.86	0.82	0.83
$\text{SKPM}_{\text{Base}}$(IC)+	0.80	0.79	0.73	0.77	0.70	0.76
RKPA-Base	0.62	0.63	0.63	0.69	0.71	0.66
Enigma+	0.61	0.69	0.58	0.55	0.69	0.64

Table 6: (YELP) Quantification precision evaluation of generated KPs. The precision is reported on five business categories: **Arts** (& Entertainment), **Auto**(motive), **Beauty** (& Spas), **Hotels**, **Rest**(aurants).

models for PAKPA leads to signficantly better performance than using an open-source LLM alone. Specifically, among multi-LLMs configurations, (`Vicuna7B + GPT3.5`) achieves state-of-the-art performance, mainly due to the powerful generative capability of `GPT3.5` on KP Generation. On the other hand, for configurations based on one LLM, (`Mistral7B + Mistral7B`) outperforms (`Vicuna7B + Vicuna7B`). It is important to note that although `Mistral7B` outperforms `Vicuna7B` on KP Generation task, `Vicuna7B` is still the top performer of the ABSA task, as shown in Table 3. This crucially contributes to the state-of-the-art performance of (`Vicuna7B + GPT3.5`).

Evaluation of KP Quantification Precision using YELP Table 6 presents the precision scores for all KPA models, which shows their general performance of matching input comments to the generated KPs across 5 business categories of YELP. Overall, PAKPA outperforms all baselines, with improvements of up to 31% in the matching precision score, and the performance is stable across the business categories. RKPA-Base, Enigma+ and $\text{SKPM}_{\text{Base}}$(IC)+, without access to the ABSA information of reviews to create aspect-specific summaries, show an inferior quantification performance compared to ABKPA and PAKPA. Integrating ABSA into the KPA system, either in extractive or abstractive techniques, then becomes a critical factor for achieving state-of-the-art performance for review summarization. For example, $\mathrm{SKPM}_{\mathrm{Base}}(\mathrm{IC})+$, whose architecture was proven effective on argument debates, achieves inferior performance when applied for reviews compared with ABKPA, an extractive KPA system incorporating ABSA. It is also worth noting that previous KPA studies with abstractive implementation, though committed to generating more concise yet less redundant KPs, always have inferior matching performance to the SOTA extractive techniques. More specifically, in most business categories, Enigma+, an early KPA system applying abstractive summarization, is outpaced by RKPA-Base, an early extractive system. Such inferiority is primarily due to data scarcity issues for fine-tuning pre-trained language models (PLM) to generate high-quality KPs for reviews, making existing abstractive KPA frameworks prone to hallucination. Interestingly, our abstractive aspect-based PAKPA system outperforms the extractive aspect-based system ABKPA, largely due to the utility of our prompted in-context learning on LLMs and aspect-oriented KP generation approach.

Error Analysis By analyzing the errors in KP generation of our system across business categories and datasets, we found several systematic patterns of errors. A frequent error type is KPs containing extra information related to its main aspects. An example KP in this category is "Overpriced breakfast with mediocre coffee". This sometimes happens when more specific aspect terms (e.g., "coffee") are clustered with more general ones (e.g. "breakfast"), and they cover different opinion information that is difficult to generalize. In some other cases, KPs generated for a cluster can also be overly generalized, and so coverage includes the major opinions of comments but may ignore the minor ones. For example, the comment "I love their pastries and they have a decent selection of yummy cookies." was matched to the aspect "Delicious and diverse cake options", which should also be referred to as the "bread" aspect.

4.4 Case Studies

We conduct case studies to evaluate the redundancy and hallucination of generated KPs for a "Hotel" business of YELP, as shown in Table 7. Overall, PAKPA stands out for generating KPs with minimal redundancy, also being highly informative and at good aspect diversity (e.g., "Poor service and unresponsive staff."), which is superior to previous abstractive counterparts such as $\mathrm{SKPM}_{\mathrm{Base}}(\mathrm{IC})+$ or Enigma+ that tend to produce repetitive, hallucinated and overly broad KPs (e.g., "didn't work at all - the front desk staff was rude, rude, and!!", "They don't listen!!!!"). Furthermore, the RKPA-Base and ABKPA models still cannot provide KPs covering sufficient aspect information and as valid and fluent as PAKPA (e.g., "Overall unprofessional and unorganized.", "are rude, slow and disrespectful."). More generated KP samples can be found in Table 10 and 11 (Appendix G).

	Key Points
PAKPA	Poor service and unresponsive staff.
$\mathrm{SKPM}_{\mathrm{Base}}$-(IC)+	didn't work at all - the front desk staff was rude, rude, and!!!
Enigma+	They don't listen!!!!
ABKPA	Overall unprofessional and unorganized.
RKPA-Base	are rude, slow and disrespectful.
CG	However, negative aspects mentioned included **issues with room conditions**, **slow service**, **noise**, **safety concerns**, and **lack of amenities**.

Table 7: KPs generated by different KPA systems summarizing a "Hotel" business of YELP

5 Conclusion

In this paper, we propose Prompted Aspect Key Point Analysis (PAKPA), a novel KPA framework applying abstractive summarization for opinion quantification. PAKPA addresses the issues of KPs with overlapping opinions, hallucination, and inaccurate quantification of previous sentence-based KPA approaches. Compared with previous studies, our approach effectively makes use of ABSA in business reviews to generate KPs grounded in aspects and achieve more accurate quantification. Experimental results show that our solution greatly enhances both the quantitative performance and quality of KPs. Secondly, our prompted in-context learning approach also deviates from the conventional supervised learning approach and removes the need for large amounts of annotated data for supervised training and fine-tuning.

Acknowledgement

This research is supported in part by the Australian Research Council Discovery Project DP200101441.

Limitations

We evaluated the textual quality of aspect KPs only on SPACE, as it is the only (to our best knowledge) public dataset with ground-truth human-annotated aspect-oriented textual summaries.

Ethics Statement

We have applied ethical research standards in our organization for data collection and processing throughout our work.

The YELP dataset used in our experiments was officially released by Yelp, while the SPACE dataset was publicly crowdsourced and released by the research publication for benchmarking opinion summarization framework. Both datasets was published by following their ethical standard, after removing all personal information. The summaries do not contain contents that are harmful to readers.

We ensured fair compensation for crowd annotators on Amazon Mechanical Turk. We setup and conducted fair payment to workers on their annotation tasks/assignments according to our organization's standards, with an estimation of the difficulty and expected time required per task based on our own experience. Especially, we also made bonus rewards to annotators who exerted high-quality annotations in their assignments.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

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

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key Point Analysis of business reviews. 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 3376–3386, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews.

Ralph Allan Bradley and Milton E. Terry. 1952. RANK ANALYSIS OF INCOMPLETE BLOCK DESIGNS: THE METHOD OF PAIRED COMPARISONS. *Biometrika*, 39(3-4):324–345.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Manav Kapadnis, Sohan Patnaik, Siba Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at ArgMining-EMNLP 2021: Leveraging pre-trained language models for key point matching. In *Proceedings of the 8th Workshop on Argument Mining*, pages 200–205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14064–14080, Toronto, Canada. Association for Computational Linguistics.

Mary McGlohon, Natalie Glance, and Zach Reiter. 2010. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 114–121.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. 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, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. *Natural language processing and text mining*, pages 9–28.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

An Tang, Xiuzhen Zhang, and Minh Dinh. 2024. Aspect-based key point analysis for quantitative summarization of reviews. In *18th Conference of the European Chapter of the Association for Computational Linguistics*.

Wenyi Tay, Xiuzhen Zhang, and Sarvnaz Karimi. 2020. Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness. *Journal of the Association for Information Science and Technology*, 71(7):784–799.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

A Prompts for `GPT3.5`

We present the zero-shot and few-shot prompts for Aspect-Based Sentiment Analysis (ABSA) and Aspect-based Key Point Generation in Listing 1 and 2.

B Details of the Experimental Datasets

SPACE A large-scale opinion summarization dataset built on TripAdvisor hotel reviews, with its test set containing a large collection of human-written summaries (for reviews of 50 hotels) usable as the ground truth in our experiment. To our best knowledge, SPACE stands out as the sole dataset providing human-written aspect-specific summaries, serving as an ideal ground truth for evaluating our aspect-based generation of KPs in PAKPA. In this experiment, we opt to select both the *general summaries*, i.e., short and high-level overview of popular opinions, and *aspect-specific summaries*, detail on individual aspects (e.g., location, cleanliness) of SPACE because they both can be represented by our KPs. Note that we ignore the aspect label of these summaries and focus only on their content in our experiment. To maintain a reasonable run time, we also limit the selection to only the top 10 hotels with the highest number of reviews in SPACE, and we exclude reviews with more than 15 sentences. We show additional statistics of our SPACE dataset in Table 8

YELP Business reviews from the Yelp Open Dataset [9], as being utilized in previous extractive KPA study for reviews (Bar-Haim et al., 2021; Tang et al., 2024), targetting five business categories; *Arts & Entertainment* (25k reviews), *Automotive* (41k reviews), *Beauty & Spas* (72k reviews), *Hotels* (8.6K reviews), and *Restaurants* (680k reviews). We applied additional filters and selections to the dataset to maintain a reasonable runtime as follows. First, we excluded reviews with more than 15 sentences. Second, on the remaining data, we target to conduct our experiment only on businesses having between 50-100 reviews, and sample for each category (e.g., hotels) the top 10 businesses with the highest number of reviews in the current filter. The process finally forms a sample of 4966 reviews (31860 review sentences) supporting 50 Yelp businesses under 5 categories to be covered in our experiment. We show additional statistics of our YELP dataset in Table 9

[9] `https://www.yelp.com/dataset`

Table 8: Statistics of SPACE

Category	# Reviews	# Sentences	# Sentences Per Review	# Sentences Per Reference Summary
Hotels	946	7510	7.94	2.48

Table 9: Statistics of YELP

Category	# Reviews	# Sentences	# Sentences Per Review
Arts	994	6000	6.04
Auto	994	6196	6.23
Beauty	995	6288	6.32
Hotels	983	7145	7.27
Rest	1000	6231	6.23

C Dimensions of KP Quality Evaluation

This section provides detailed descriptions of tasks and dimensions involved in our manual evaluation of the KP textual quality. Annotators were asked to perform a pairwise comparison between two sets of KPs, each taken from a different model, generated for a specific reviewed business entity considering a specific dimension. The annotators must answer a comparative question with respect to the evaluating dimension. (e.g., *Which of the two summaries captures better ...*). For each dimension, following Friedman et al. (2021), we calculate the ranking using the Bradley-Terry model (Bradley and Terry, 1952), which predicts the probability of a given participant winning a paired comparison, based on previous paired comparison results of multiple participants, and thus allows ranking them.

Listing 1: Few-shot prompt (18 examples) for prompting `GPT3.5` on fine-grained Aspect-based sentiment analysis. Please refer to our released code for full prompts.

```
You will be provided with a review sentence delimited by triple quotes.
A review sentence usually covers the customer opinions expressed on different aspects of a product or service.

You were tasked to perform Aspect−based Sentiment Analysis to extract the user sentiments expressed on different aspects in the review.
Formally, we define subtask of extracting the aspects it corresponding sentiments as Aspect Extraction and Aspect Sentiment Classification:
− Aspect Extraction: Identifying aspect targets in opinionated text, i.e., in detecting the specific aspects of a product or service the opinion holder is either praising or complaining about. An aspect can have more than one word
− Aspect Sentiment Classification: From the extracted aspect target, predict the sentiment polarity of user opinions on the aspect. The sentiment polarity value can be: "positive", "neutral", and "negative".

Provide the answer in JSON format with the following keys: aspect, sentiment

Review sentence: \"\"\"Movies cost $ 14 , and there is no student discount at this location .\"\"\"
Answer: [{'aspect': 'student discount', 'sentiment': 'negative'}]

Review sentence: \"\"\"Our tour guide was knowledgeable about the property and about all things Frank Lloyd Wright .\"\"\"
Answer: [{'aspect': 'tour guide', 'sentiment': 'positive'}]

Review sentence: \"\"\"BMW Henderson made my purchase easy and stress free .\"\"\"
Answer: [{'aspect': 'purchase', 'sentiment': 'positive'}]

Review sentence: \"\"\"I had a male therapist and he was amazing !\"\"\"
Answer: [{'aspect': 'male therapist', 'sentiment': 'positive'}]

...

Review sentence: \"\"\"Be sure to accompany your food with one of their fresh juice concoctions .\"\"\"
Answer: [{'aspect': 'food', 'sentiment': 'neutral'}, {'aspect': 'fresh juice concoctions', 'sentiment': 'positive'}]

Review sentence: \"\"\"During busy hrs, i recommend that you make a reservation .\"\"\"
Answer: [{'aspect': 'reservation', 'sentiment': 'neutral'}]

Review sentence: \"\"\"The menu, which changes seasonally, shows both regional and international influences .\"\"\"
Answer: [{'aspect': 'menu', 'sentiment': 'neutral'}]

Review sentence: \"\"\"Our waitress had apparently never tried any of the food, and there was no one to recommend any wine .\"\"\"
Answer: [{'aspect': 'waitress', 'sentiment': 'negative'}, {'aspect': 'food', 'sentiment': 'neutral'}, {'aspect': 'wine', 'sentiment': 'neutral'}]
"""
```

Listing 2: One-shot prompt for prompting GPT3.5 on KP Generation.

```
You will be provided with a list of user review comments delimited by triple quotes, and a list of common aspects shared by
        those reviews delimited by triple quotes
The comments in the list has been clustered by some common aspects and sentiment.
You were guided to generate a concise key point that captures opinions on the most popular aspect across the input comments,
        and also accomodate the provided list of common aspect.
Note that the generated key points must describe the opinion in only ONE aspect only and must not discuss multiple aspects.
        The generated key points must have 3-5 tokens.

Perform the following actions to solve this task:
- Identify the single and general aspect (e.g. atmosphere) that are common across the input aspects terms
- On the identified aspect, find the salient points of opinions mentioning that aspect across the input comments
Some invalid examples of key points with multiple aspects that must be avoided:
- "Enjoyable atmosphere with great music and live entertainment.", rather it should be "The atmosphere is very enjoyable."
- "Excellent wine selection and enjoyable atmosphere.", rather it should be "The wine selection is great."

Comments: """['The bartenders were so sweet and were very responsive .', 'The staff is fantastic and responsive .', 'The staff
        was so accommodating and kind !', 'The hotel staff went above and beyond with their customer service .', 'The staff was
        super accommodating and made planning a cinch .', 'Front desk staff was welcoming and accommodating .', 'All staff
        were friendly , helpful & professional . ', 'Everyone of the staff has been super friendly and accommodating .', 'Rooms
        are comfortable and staff are friendly .', 'The staff was courteous & informative .', 'Mandatory valet parking with
        excellently quick service and attentive desk staff .', 'Much better location and competent staff !', 'The staff is amazing -
        upbeat , involved , and made great recommendations . ', 'The front desk staff was unbelievably friendly and
        accommodating .', 'Clean , comfortable and friendly , accommodating staff .', 'Their service was professional ,
        accommodating , fast and cordial .', 'The staff was friendly and rectified any mistakes on our reservation .', 'The front
        staff is accommodating , informative , and friendly . ', 'The staff was courteous and efficient .', 'The staff was friendly
        and courteous .', 'Pool , spa , gym -- super courteous staff , what more could you want ?']"""
Aspects: """['bartenders', 'staff', 'hotel staff', 'front desk staff', 'desk staff', 'front staff']"""
Key Point: Friendly and helpful staff .
```

- VALIDITY: The key point should be an understandable, well-written sentence representing an opinion of the users towards an aspect of the business entity. This would filter out sentences such as *"It's rare these days to find that!"*.
- SENTIMENT: The key point should have a clear sentiment towards the business entity under reviewed. (either positive or negative). This would exclude sentences like *"I came for a company event"*.
- INFORMATIVENESS: It should discuss some aspects of the reviewed business and be general enough. Any key point that is too specific or only expresses sentiment cannot be considered a good candidate. Statements such as *"Love this place"* or *"We were very disappointed"*, which merely express an overall sentiment, should be discarded, as this information is already conveyed in the star rating. The KP should also be general enough to be relevant for other businesses in the domain. A common example of sentences that are too specific is mentioning the business name or a person's name (*"Byron at the front desk is the best!"*).
- SINGLEASPECT: It should not discuss multiple aspects (e.g., *"Decent price, respectable portions, good flavor"*).
- REDUNDANT: Each KP should express a distinct aspect. In other words, there should be no overlap between the key points.
- COVERAGE: A set of KPs should cover a wide diversity of opinions relevant and representative of the reviewed business.
- FAITHFULNESS: KPs should express reasonable and meaningful opinions to the reviewed business without hallucination. No conjecture or unfounded claims arise.

D Pairwise KP Quality Comparison Annotation Guidelines

Below are the two summaries for a business in *Arts & Entertainment*, generated by two different summarization frameworks. Each summary contains several key points (i.e., salient points) generated summarizing the user opinions on different aspects. You are tasked to select which summary you think is better according to the below criteria.

Business: Saenger Theatre.

Criteria: REDUNDANCY. Each key point in the summary should express a distinct aspect. In other words, there should be no overlap between the key points.

Summary A: ['The Saenger Theater is a beautiful and stunning venue.', 'Comfortable seating.', 'Great shows.', 'Beautiful and impressive renovation.', 'Excellent acoustics and sound quality.', 'Technical issues during the performance.', 'Limited and uncomfortable bathroom space.', 'Show cancellations and disruptions.', 'Uncomfortable seats and high seat prices.', 'Disappointing theater experience.']

Summary B: ['The renovations of the theater were praised.', 'The theater had exceptional shows.', 'Canceled shows were criticized.', 'The venue is stunning.', 'The staff at the theater was great.', 'Limited space in the bathroom was criticized.', 'The setup of the bathrooms was odd.', "The theater's location received negative comments."]

The options are:

- Summary A
- Summary B

E Key Point Matching Annotation Guidelines

Below are the match annotation guidelines for (sentence, KP) pairs:

In this task you are presented with a business domain, a sentence taken from a review of a business in that domain and a key point.

You will be asked to answer the following question: does the key point match the sentence?

A key point matches a sentence if it captures the gist of the sentence, or is directly supported by a point made in the sentence.

The options are:

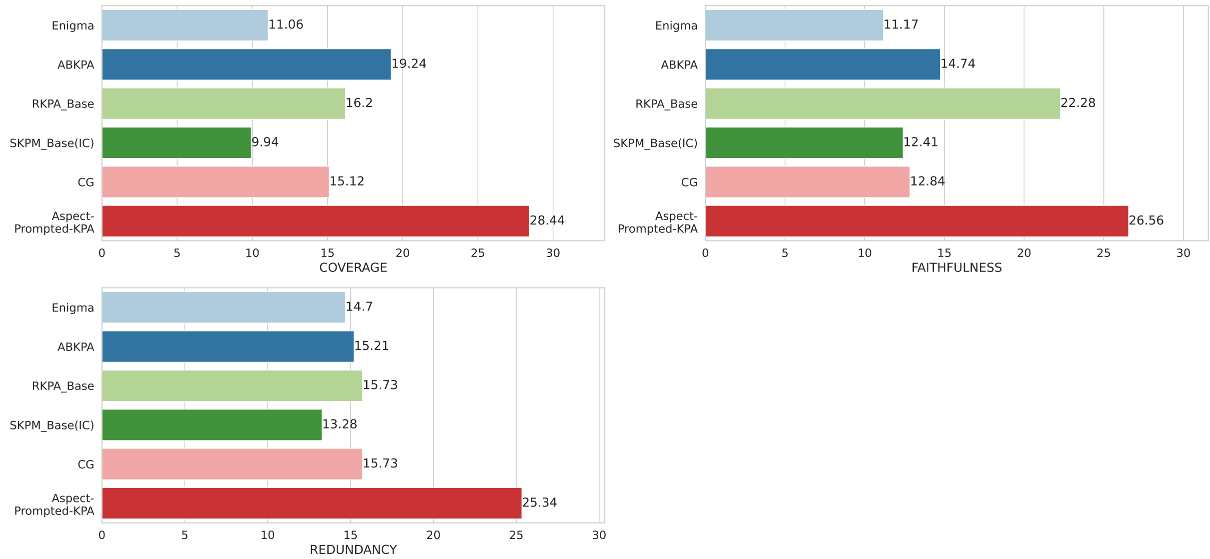
- Yes
- No
- Faulty key point (not a valid sentence or unclear)

F Comparative Analysis of KP Quality: A Visual Overview

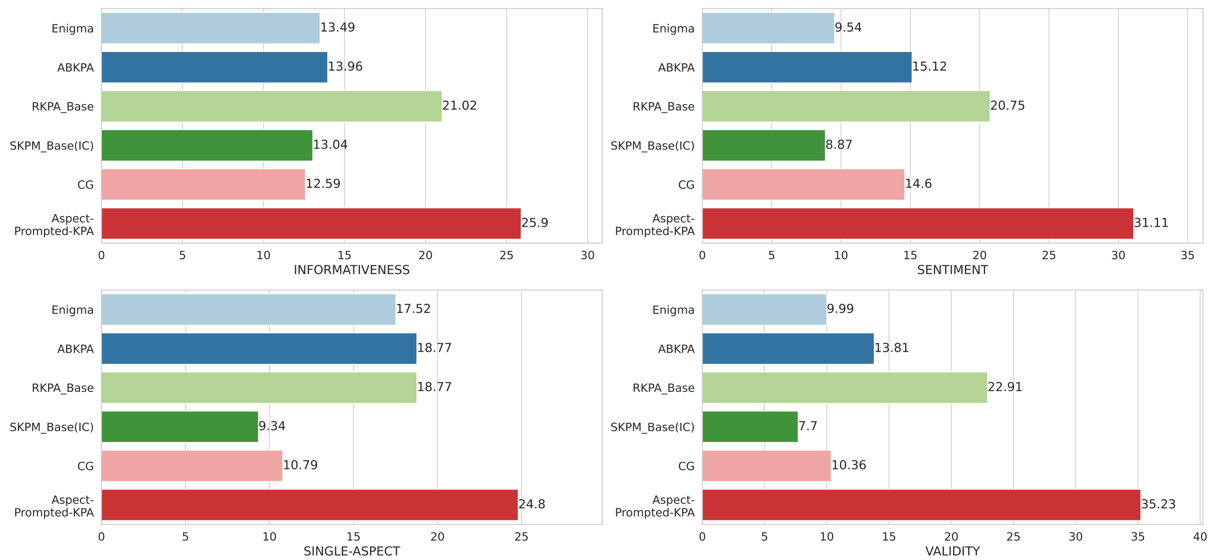
Figure 2 visualizes the Bradley Terry scores. as already presented in Table 5, in bar charts for more comprehensive view of our human evaluation results on different KPA systems.

G Summary of KPA Frameworks and Prompted Opinion Summarization Framework

This section presents details of Table 10, which shows some top negative KPs for all KPA systems, ranked by their prevalence and compares with the textual summary generated by the traditional prompted summarization framework (using GPT3.5) (CG).



(a) KPs as summaries for salient points from corpus



(b) KPs as summaries for reviews

Figure 2: Bradley Terry scores of comparative human evaluation of different KPA frameworks on 7 dimensions in assessing how well they summarize the corpus (2a) and provide KPs for reviews (2b).

PAKPA	**SKPM$_{\textbf{Base}}$(IC)+**	**Enigma+**	**ABKPA**	**RKPA-Base**
Issues with the room and front desk service.	didn't work at all - the front desk staff was rude, rude, and!!!	They don't listen!!!!	Cons:* Very noisy rooms.	Overall unprofessional and unorganized.
Terrible hotel experience.	didn't have a receptionist at the front desk.!!!	Called front desk.	Overall unprofessional and unorganized.	Carpet was stained and filthy.
Difficult and expensive parking options.	a hotel is a "non smoking" hotel.!!!	They did not plan ahead!	And parking was also overpriced.	It didn't feel safe.
Poor service and unresponsive staff.	I would never stay here again.!!!	Hotel is disgusting.	Poor hotel for the price.	are rude, slow and disrespectful.
Issues with shower and bathroom cleanliness.	was a bit of a walk from the hotel to the parking lot.!!!	Would not recommend this hotel.	The food service was slow.	beds are very lumpy.
. . .				
Recursive GPT-3-Chunking (CG): . . . . However, negative aspects mentioned included **issues with room conditions**, **slow service**, **noise**, **safety concerns**, and **lack of amenities**. . . .				

Table 10: Top 5 negative-sentiment key points, produced by experimenting KPA systems, ranked by their prevalence on a "Hotel" business on YELP, comparing with the textual summary created by the prompted opinion summarization framework (CG).

PAKPA	**SKPM$_{\textbf{Base}}$(IC)+**	**Enigma+**	**ABKPA**	**RKPA-Base**
Excellent bakery with delicious treats.	has a good selection of pastries, pastries, pastries, and pastries!!!	Bread, baguettes, fresh.	Love love love this place.	Great baked sweets and breads.
Delicious and diverse cake options.	has a good selection of pastries/cookies/-cookies/c!!!	The best bread in Tucson.	Cappuccino and croissants are delish!	Prices are extremely reasonable!
Friendly and efficient staff.	Sprouts' has a good selection of breads and pastries.!!!	You gotta go here!!!	Clean and well staffed.	They're worth the wait!
Excellent prices.	Definitely recommend this place to anyone looking for a good!!!	The food is delicious.	Great baked sweets and breads.	Great food and flavor!
Delicious baked goods.	I will definitely be back.!!!	Very friendly staff.	Prices are extremely reasonable!	Best friendly service, ever!
Irresistible smells and incredible taste.	has the best bread in Tucson at a reasonable price.!!!	It was delicious!	Always hot and fresh tasting.	Familiar yet unique!
Enchanting and beloved place.	I've been to this bakery for 20 years!!!	Nice old school bakery.	Great stop for lunch.	Amazing food and friendly service.
...				
Recursive GPT-3-Chunking (CG): ... The bakery is highly regarded as the best in Tucson, with high-quality products. ... Specific items like the baguette, sesame rolls, and dinner roll were highly rated for their taste, texture, and reasonable prices. ... Customers appreciated the bakerys "old school" vibe, excellent prices, and consistently wonderful French bread and pastries. ... Customers also praised the early opening hours, friendly staff, and variety of baked goods available. ...				

Table 11: Top 7 positive-sentiment key points, produced by experimenting KPA systems, ranked by their prevalence on a "Restaurant" business on YELP, comparing with the textual summary created by the prompted opinion summarization framework (CG).