A Viewpoints Embedded *Diff-table* System For Cross-sectional Insight Survey In a Research Task

Jinghong Li[1], **Naoya Inoue**[1], **Shinobu Hasegawa**[1]
Japan Advanced Institute of Science and Technology, Japan

Abstract

In the flourishing era of information science, effective comprehension, observation, and insight from various academic papers are crucial skills for researchers. However, this can be challenging for beginners without enough research training. The current knowledge graphs and automatic summarization systems used in research insight surveys rarely highlight the similarities and differences among multiple papers based on agreed-upon expert features. This can make novice researchers difficult to understand the logical connections between research concepts. Therefore, this study is committed to assisting researchers in conducting Cross-sectional Insight Survey. It offers a concise ***diff-table*** output format, tailored from the perspective of expert consensus. This study aims to generate a table of abstractive summarization based on the viewpoints of expert consensus and showing the differences under these consensus. The final output is in the form of a concise ***diff-table*** to assist researchers in conducting Cross-sectional Insight Survey. Our evaluation demonstrates that our generated ***diff-table*** outperforms the baseline in terms of *BERTScore* and conciseness.

1 Introduction

With the advancement of information science, the number of academic papers has increased exponentially. Consequently, it is crucial to quickly understand the research concepts, the underlying logic, and the task dynamics of specific fields from such a vast and continuously growing database for research surveys (Altmami and Menai, 2022; Li et al., 2024a, 2023a). Li et al. mainly assisted novice researchers from two perspectives in conducting their research surveys more efficiently: **(1)** *the bird's eye view survey*, which determines the causal logic in research issue (Li et al., 2024c), and **(2)** *the insight survey*, which analyzes the relevance and inheritance among articles (Li et al., 2024b).

Both of them rely on the issue ontology extracted from the 'introduction' and 'conclusion' sections. These issue ontologies are used to classify sentences and generate knowledge graphs based on their summarization output. These two methods facilitate longitudinal survey (Cook et al., 2002), allowing for cause-and-effect comparisons across multiple papers, and enabling researchers to track changes and patterns during a specific period. However, relying solely on the longitudinal survey via issue ontology set-based lacks in-depth analysis of the research content, which is drawn from the consensus views of experts in the research field such as datasets, pre-training model experts used, performance experts achieved, etc. which often appear in the Natural Language Processing (***NLP***) research field. Considering this expert consensus, it is clear that authors often produce similar content from certain viewpoints. They also express unique aspects based on these viewpoints, reflecting their research originality and differentiating their work from others. Therefore, it is important for novice researchers to understand and compare content cross-sectionally via expert consensus from research tasks, to identify unique, high-impact characteristics for executing an in-depth insight survey.

One way to support the Cross-sectional insight survey is using prompt engineering based on *ChatGPT* to generate abstractive summarization (Luo et al., 2023; Velásquez-Henao et al., 2023). Viewpoints can also be embedded as column header to generate table reflect differences (***diff-table***) from multiple articles. However, our experiments will show that over-reliance on *ChatGPT* without proper prompt description and input text does not produce satisfactory ***diff-table*** because of two reasons. First, if the input data are not properly pre-processed, irrelevant information may interfere with the output accuracy, especially when dealing with large text inputs that have a high number of useless tokens for summarization. Second, *Chat-*

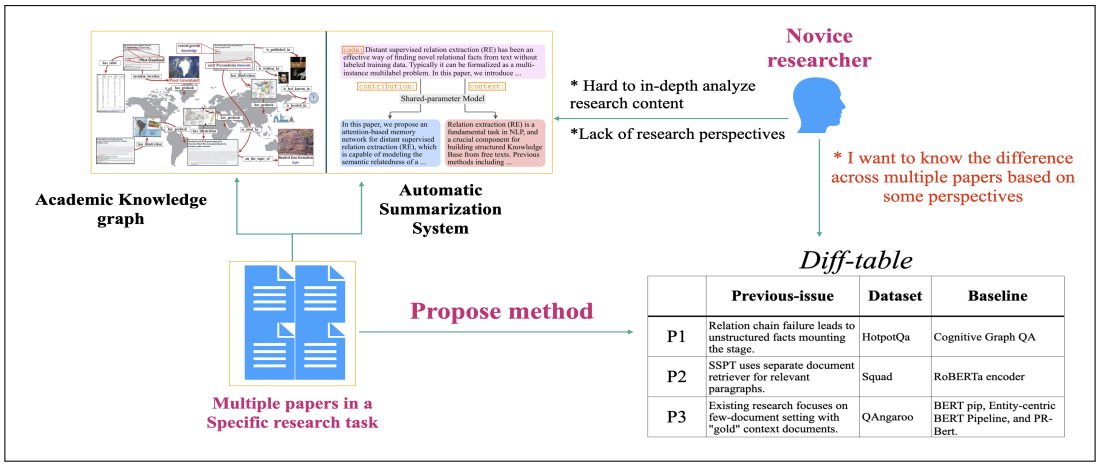

Figure 1: The feature of diff-table, different from academic knowledge graph (Deng et al., 2021) and automatic summarization system (Hayashi et al., 2023).

GPT's lack of professional research training can make it difficult to locate original texts that reflect expert consensus in the research field. This could result in issues with the incomprehensibleness and completeness of the generated summary (Dönmez et al., 2023; Rahman et al., 2023).

To address the above issues, this study aims to develop a system that assists researchers in the Cross-sectional Research Insight Survey through abstractive summarization in a viewpoints-embedded ***diff-table*** format. As shown in Figure 1, unlike previous systems, our ***diff-table*** consists of abstractive summarization cells and helps researchers identify similarities, unique aspects, and impacts of the research task, enabling a more efficient insight survey. Experimental results indicate that our tool outperforms existing support tools based on *ChatGPT* + prompt engineering in terms of both information accuracy and conciseness, showing potential for further development. Our main contributions are as follows.

1. A ***diff-table*** system for Cross-sectional Insight Research Survey. We specially develop a dataset based on *S2orc* (Lo et al., 2020) for this purpose and use this dataset to automatically generate the ***diff-table***.

2. Viewpoints-embedded template in ChatGPT prompts, which are used to generate an abstractive summarization for each cell in the ***diff-table***.

2 Related work

Supporting the Cross-sectional Insight Survey involves condensing information from various academic papers and highlighting their commonalities and differences. Automatic summarization is one method that can be used to achieve this, as it provides a concise output to make it easier for novice researchers to understand the research content quickly. However, recently, most automatic summarization or knowledge graph support systems have tended to favor longitudinal surveys. For example, they track developments from ancient times to now, identify shifts in user interests and capture their evolution through time (McKeown and Radev, 1995; Vassiliou et al., 2023; Zhang et al., 2024) or excavate the inheritance relationship of the paper itself (Li et al., 2024b). The summary generated in this way may not include consensus views from the research field, making it difficult to compare differences among multiple articles with a similar research task. Furthermore, knowledge graphs such as (Ammar et al., 2018; Chen and Luo, 2019; Xu et al., 2020), consisting of academic papers with numerous articles, are primarily made up of citation relationships and keywords in that research field. The representation of these summary may often be high-dimensional, which may overwhelm novice researchers due to the complexity in understanding the knowledge logic.

On the other hand, the method that embeds viewpoints, such as emphasizing the context of 'contribution' or 'limitation' of the article, provides insight into the research direction (Hayashi et al., 2023; Liu et al., 2023; Chen et al., 2022; Faizullah et al., 2024). However, it is not easy to discern the main purpose of the research paper solely from the content of the contribution context, because it is impossible to derive additional comparative viewpoints to highlight differences among multiple papers from that purpose.

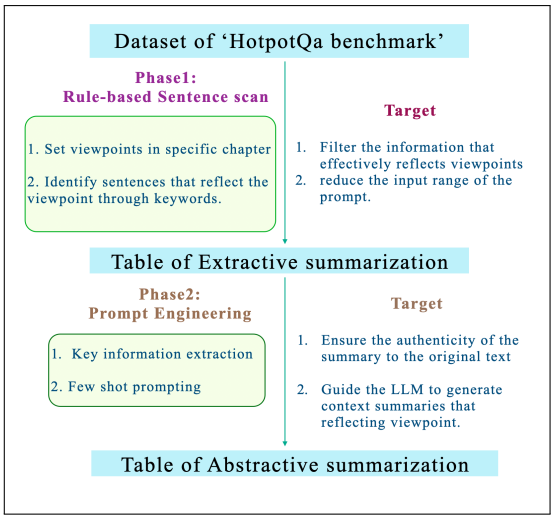

Figure 2: Overview of ***diff-table*** system development

To address the limitations, we propose a ***diff-table*** output form. This format can express the differences within each study, under the consensus of field experts. This tool makes it easier for researchers to compare the commonalities and differences across numerous articles, providing a unique guidance for novice researchers.

3 Methodology

We begin by defining viewpoints, Cross-sectional Insights, and ***diff-tables***. Then, we sequentially describe the process of generating ***diff-tables*** as detailed in Figure 2. We focus on the content of academic papers in a specific research task as input text of system. Our primary strategy involves performing extractive summarization first to narrow down the input text of ***LLM***, aiming to reduce the impact of text that is not related to the specified viewpoint. We then take this condensed text and use it for prompt engineering, generating abstractive summarization and ***diff-table***. The prompt we crafted maintains the integrity of the original content, while attempting to cover the important information that reflects specific viewpoints.

3.1 Definition

3.1.1 Viewpoints in research field

Viewpoints refer to the research methods agreed upon among experts in a given research task. This consensus has been gathered from the inception of the research area to the present day, forming a unified viewpoint (Li et al., 2023b). Most of the papers in a research task are structured around specific viewpoints. Therefore, it is essential for novice researchers to understand and use these viewpoints to discover key points in their research activity.

3.1.2 Cross-sectional Insight Survey

Cross-sectional study aims to identify differences between groups, helping researchers understand various situations at certain time (Wang and Cheng, 2020). In this study, we expand our focus to a Cross-sectional Insight Survey on research tasks. This survey style outlines the fundamental attributes of the research task and expresses the difference under these attributes. The advantage of this method is that the indicators are typically unified on the basis of experts' consensus. Deep-mining this consensus, some commonalities and differences could be discovered in each article. This approach of identifying differences through consensus offers researchers a perspective for in-depth analysis of research topics and key information.

3.1.3 Diff-table

The ***diff-table*** is an output format of the Cross-sectional Insight Survey. It organizes data based on differing viewpoints. This table includes summary cells from various articles, with the viewpoints represented as column headers. For example, in this research, the viewpoints we define refer to the consensus of experts in the field of ***NLP***, as shown in Table 1. The abstractive summarization of the paper is consolidated into cells that reflect specific viewpoints. This ***diff-table*** format facilitates the comparison of similarities and differences among papers, assisting in the analysis and comprehension of various research elements (Chen, 2023). In this work, ***diff*** means difference that refers to the distinctions of the summaries in multiple cells.

3.2 Extractive Summarization based on viewpoints

This section introduces the extractive summarization process of papers to limit the text input scope to the ***LLM***. We first use the two-stage semantic text matching (McKeown and Radev, 1995; Zhong et al., 2020) method of paper → paragraph → sentence to extract key sentences that reflect the viewpoint. Content reflecting a particular viewpoint typically appears in specific sections of an article and includes certain keywords [1]. For instance, previous-issue usually found in the introduction

[1] https://fastercapital.com/content/Effortlessly-summarize-articles-with-best-summary-generator.html

Table 1: Configuration of extractive summarization reflect viewpoints

Viewpoint	Keyword	Section range	Definition
Previous issue	- however - difficulty, limit	- Introduction - Related work	Unresolved problems in Previous Research mentioned in this article
Objective	- we propose - in this study - we aim	- Introduction - Related work - Conclusion	The main propose of this article
Dataset	- we/our + dataset	- Except Introduction and Related work	The dataset mainly used or developed in this article
Pre-training model	- we/our + pre-train	- Except Introduction and Related work	The pre-training model mainly used or developed in this article
Baseline	- baseline	- All	The strategy of setting the baseline to execute experiment
Performance	- we/our + performance - achieve, outperform	- All	The work carried out by the authors and the performance they obtained
Limitation	- limitation	- Limitation - Case study - Conclusion	The authors point out the limitations of their proposed method.
Future work	- future - further	- Limitation - Case-study - Conclusion	The future directions mentioned by the authors

and related work sections, often start with the keyword "however". Thus, to create an abstractive summary that accurately captures these viewpoints, we first need to perform extractive summarization. This process determines the text input range for the abstractive summarization stage. To execute an extractive summarization, we first need to identify sentences that contain viewpoint features in the paper. This process begins by locating the specified section to narrow down the search range. Next, we scan the paragraphs within this range, identifying sentences that include viewpoint keywords for extraction. We extract not only the sentences expressing the viewpoint but also the preceding and following sentences to accommodate key information that appears in their context. One criterion we set is that the sentences should reflect the article author's unique descriptions for each viewpoint, rather than descriptions of related studies. We determine keywords for each viewpoint based on the prevalent features of ***HotpotQA benchmark task***, as depicted in Table 1. This extractive summarization contains both viewpoint information and non-viewpoint information, which needs to be further screened and summarized by the next step of prompt engineering.

3.3 Abstractive summarization in *diff-table*

We use the prompt engineering via ***LLM*** - *gpt-4o-mini* [2] model to generate abstractive summarization for each cell, using the extractive summarization as input. This process is divided into two stages. The first stage involves extracting only the relevant viewpoint information from each sentence and filtering out any unimportant information that does not affect the reading. Although this stage outputs a simplified summary, there may be some repeated information in multiple sentences. Hence, in the second stage, we further compress the output summary of the first stage for each cell by organizing repeated information to further condense the summary.

3.3.1 Prompt-engineering: Viewpoint Refinement

In the initial stage of prompt-engineering, our goal is to identify important information that reflects the viewpoint within sentence chunks. The comprehensiveness of the summary output depends on the description of the prompt. To guide the ***LLM*** generates precise and concise summaries, follow these three points:

1. Precisely retain the essential information from the original text.

2. Eliminate content that does not reflect any viewpoints and does not affect readability.

3. Prevent the ***LLM*** from generating tokens that contradicts the facts of original text.

Using the Zero-shot method without guiding the output can lead to verbose summaries or summaries lacking key information. To enhance this, we adopt the Few-shot method (Zhang et al., 2022), incorporating an example into each prompt description to guide the model towards context imitation. Table 5 presents an example of each viewpoint summary.

The sample description of prompt in the informa-

[2] https://platform.openai.com/docs/models/gpt-4o

tion identification stage is shown below: The settings of the three variables, **eg_org** (sample of original text), **eg_output** (sample of summary based on original text), and **kp** (feature of viewpoint refer to Table 5).

```
prompt = f""" Your task is to extract relevant information from text to make a brief summary in a consistent style.
    <Original text>:{eg_org}

    <Summary>:{eg_output}

    From the original text below, delimited by triple quotes, extract the information only relevant to {kp}. Try to decrease the usage of adjectives and adverbs for a more concise summary. If no relevant information is found, do not output.

    <Original text>: ```{text}```
     """
```

Listing 1: Prompt: Viewpoint-text Identification

3.3.2 Prompt-engineering: Compression

After the initial stage of prompt-engineering, some cell of summaries may contain repetitive content. This happens when the same viewpoint is extracted from different chunks multiple times. For example, if an article mentions the *HotpotQa dataset* in several sections, our focus is solely on the datasets used in the article. These summaries require further refinement to streamline repetitive and wordy segments. To reduce verbosity, the second stage of prompt-engineering is mainly focused on identifying and removing redundant information without negatively impacting the tokens in summary. Here is a sample detailed explanation of the process.

```
prompt = f""" Your task is to compress text in a consistent style.
    <Original text>: HotpotQA, HotpotQA, full wiki opendomain QA setting, opendomain QA datasets, opendomain QA datasets, HotpotQA dataset

    <Compressed text>: HotpotQA dataset, full wiki opendomain QA setting, opendomain QA datasets

    Please compress the following text, delete repetitive expression without altering the meaning.

    <Original text>: ```{text}```
     """
```

Listing 2: Prompt: Compression

4 Diff-table Evaluation

We conducted the evaluation experiment for ***diff-table*** in three stages. First, we manually created the gold standard of ***diff-table*** for 18 articles from the Papers with Code website. Next, we used *BERTScore* to objectively evaluate and compare the abstractive summarization in ***diff-table***. Lastly, we subjectively evaluate of ***diff-table*** from four perspectives: Consistency, Correctness of Viewpoint ***(VP)***, Comprehensible, and Sufficient Coverage ***(SC)*** to validate the effectiveness of ***diff-table***.

4.1 Data-processing

This study uses data from the ***HotpotQa benchmark task*** (Yang et al., 2018), as listed on the Papers with Code website[3]. The paper's title is extracted from this page using web scripting, which allows us to match the data of the original academic paper from *S2orc* dataset[4] - a corpus of 81.1 million academic papers in English (Lo et al., 2020). The corresponding papers' text and section annotation are then extracted to serve as the system's input data. Subsequently, based on these input data, both extractive and abstractive summarizations are generated via our ***diff-table*** system.

4.2 Gold standard

To objectively and subjectively evaluate the performance of the generated ***diff-table***, we reviewed the target articles and established a gold standard, following the writing standards based on the definition of viewpoint in Table 1 and the output features (summary style) in Table 5. While creating the Gold standard, we focus on the following aspects:
1. Concentrate on the facts, considering their specific characteristics, and ignore the part of the analysis and the detailed explanation.
2. If an input text represents multiple viewpoints, summarize only the content of the specific viewpoint, ensuring there is no overlap with the summary of another viewpoint.

4.3 Objective evaluation

To objectively evaluate the generated ***diff-table***, we use *BERTScore* (Zhang et al., 2019) to compare each cell of the ***diff-table*** with the gold standard, assessing the correctness of the generated abstractive summarization. We objectively compare its performance with similar ***diff-table*** generation tools,

[3] https://paperswithcode.com/sota/question-answering-on-hotpotqa
[4] https://github.com/allenai/s2orc

Table 2: Evaluation of abstrative summrization - **Left: *BERTScore*(Average F_1) | Right: Redundancy rate**
Scispace (No *VP* description): Prompt engineering with solely viewpoint names as input.
Scispace (Included *VP* description): Embed the names and description of viewpoints into prompt engineering.

	Our approach (Zero-shot)	**Our approach (Few-shot)**	**Scispace (No *VP* description)**	**Scispace (Include *VP* description)**
Previous-issue	0.67 / 1.56	**0.71 / 0.98**	0.61 / 1.99	0.61 / 1.7
Objective	0.68 / 1.81	**0.72 / 1.23**	0.65 / 3.96	0.68 / 2.2
Dataset	0.66 / 1.58	**0.68 / 1.18**	0.58 / 11.37	0.57 / 9.29
Pre-training	0.65 / 0.61	**0.66 / 0.5**	0.55 / 5.49	0.57 / 2.45
Baseline	**0.66 / 0.6**	**0.66** / 0.76	0.57 / 6.13	0.57 / 5.98
Performance	0.64 / 1.64	**0.68 / 1.52**	0.64 / 1.64	0.65 / 1.72
Limitation	**0.67** / 1.04	**0.67 / 0.99**	0.58 / 5.33	0.61 / 3.04
Future-work	0.67 / 1.3	**0.7 / 0.8**	0.65 / 4.09	**0.7** / 2.26

such as *Scispace*[5]. Unlike the traditional n-gram evaluation method that relies on original tokens, *BERTScore* computes a similarity score for each token in the candidate sentence against each token in the reference sentence. Since the tokens generated by the AI may not always be based on the original text, employing *BERTScore* to evaluate our ***diff-table*** could serve as a more fitting indicator. We select the *scibert_scivocab_cased*[6] pre-training model, which was trained using a corpus of scientific papers, as the evaluation model for *BERTScore* (Beltagy et al., 2019). This training corpus consisted of papers from Semantic Scholar. The size of the corpus was 1.14 million papers with 3.1 billion tokens included in the full text used for training. *scibert_scivocab_cased* exhibits adaptability to both the corpus and domain, making it suitable for our objective evaluation. The accuracy of the summary of each viewpoint is determined by averaging the F_1 of *BERTScore* across 18 articles. In the column where each viewpoint is located, calculate the average *BERTScore* for all cells in that column and exclude any cell without a corresponding viewpoint summary from the *BERTScore* calculation. Furthermore, the conciseness of the summary is evaluated by comparing the length of the generated summary with the gold standard expressed as redundancy rate, calculated by the ratio of the length of the generated text strings to the length of gold standard strings. The higher the value of the redundancy rate, the more redundant information included in the summary.

The evaluation results are shown in Table 2. It becomes apparent that Few-shot outperforms Zero-shot methods in both the *BERTScore* score and the level of abstract compression. Additionally, it exceeds *Scispace*'s prompt engineering **(Collect on the day of 2024/08/18)** in most aspects. This improvement of performance can be attributed to our strategy of controlling the input text range from extractive summarization, and our prompt description with viewpoint refinement style. Meanwhile, in most cases, the summaries generated by the Few-shot method are more concise than those produced by the *Scispace* and Zero-shot methods, Proves that Few-shot method can more effectively remove redundant information and perform more closely approach to the gold standard.

Next, we conduct a subjective analysis of the ***diff-table*** table for several aspects. For comparative analysis with *Scispace*, we employ their more effective 'include viewpoint description' prompt to carry out our experiments.

4.4 Subjective evaluation

While ***LLM*** may sometimes generate expressions similar to the original text, these expressions may lack precision for academic fields and can lead to ambiguity. There is also a minor risk that the generated summary might modify certain proper nouns. Hence, solely using *BERTScore* evaluation is not sufficient to accurately measure the effectiveness of the summary. One case study illustrates that compared to the gold standard shown in Table 4, the Few-shot method, while removing some subjects and adjectives to shorten the summary, may also eliminate useful information to understand the content. In contrast, the Zero-shot method, due to its lack of summary examples, adds non-essential expressions that do not impact comprehension. Additionally, without a clear limit on text input, *Scispace* and ***LLM*** may struggle to select important information that reflects the viewpoint, often resulting in relatively lengthy summaries. This type of case is difficult to evaluate solely using *BERTScore*. Thus, it is necessary to adopt a method for human

[5] https://typeset.io
[6] https://huggingface.co/allenai/scibert_scivocab_cased

assessment of the summary's quality. To improve the shortage of objective evaluation, we refer to the definition of (Inoue et al., 2021; Aharoni et al., 2023) to adopt subjective evaluation methods compared to the gold standard to measure the effectiveness in four aspects:

1. Consistency: The factual consistency between the summary and the original source (input text of the prompt) (Fabbri et al., 2021)

2. Correctness of *VP*: Whether the summary content containing viewpoints is correct.

3. Comprehensible: The expression of viewpoint reflection, whether the reader can understand the general meaning of the sentences and find the key-points of the survey that directly reflect the viewpoint.

4. Sufficient Coverage *(SC)*: whether the important information that directly reflects the viewpoints of the sentence has been fully expressed. In subjective evaluation, we should initially concentrate on the correctness and comprehensibility of the summary because we can only evaluate sufficient coverage if the generated summary is correct.

Based on the four aspects outlined above, we establish the following scoring step.

1. <1> In comparison to the gold standard, a generated summary earns a score of ***+2*** if it contains sentences that are consistent, express correct viewpoints, and are comprehensible. **<2>** If the summary matches the criteria for consistency and Correctness of *VP*, but lacks readability (either too verbose or too concise), the score will be ***+1***. **<3>** If more than 50% of the entries in the summary cell are either too verbose or too concise, it is considered poorly comprehensible and receives a score of ***0***. **<4>** If the summary's content contradicts the facts in the original text, it will receive a ***-2*** points penalty. **<5>** Summary that only include incorrect viewpoints receives a score of ***-1***.

2. The second stage evaluates the degree of sufficient coverage of the correct sentences in relation to the gold standard. This involves calculating the ratio of sentences in a cell that align with the consistency of the gold standard sentence, as demonstrated:

$$SC = \frac{Count_{fully\_expressed}}{Count_{GD}} \quad (1)$$

$Count_{fully\_expressed}$: The number of sentences in the summary that fully expressed the gold standard sentence

$Count_{GD}$: The number of sentences in the gold standard cell.

Table 3: Subjective Evaluation - The average score of 18 articles for each viewpoint: Consistency & Correctness of ***VP*** & Comprehensible (**C**), Sufficient Coverage (**SC**)

	Zero-shot	**Few-shot**	**Scispace**
Previous issue	C : 1.40 SC : 0.74	**C : 1.56** **SC : 0.83**	C : 0.33 SC : 0.42
Objective	C : 1.22 SC : 0.75	**C : 1.40** **SC : 0.78**	C : 1.27 SC : 0.70
Dataset	C : 0.36 SC : 0.64	C : 0.39 SC : 0.61	**C : 0.44** **SC : 0.70**
Pre-training	C : 0.17 SC : 0.54	C : 0.17 SC : 0.58	**C : 0.26** **SC : 0.68**
Baseline	C : **0.93** **SC : 0.61**	C : 0.86 SC : 0.60	C : 0.75 SC : 0.52
Performance	C :1.11 **SC : 0.67**	**C : 1.33** **SC : 0.67**	C :1.27 SC : 0.60
Limitation	C : 0.55 SC : 0.41	**C : 0.80** **SC : 0.45**	C : -0.25 SC : 0.32
Future work	C : 0.86 SC : 0.55	**C : 1.14** **SC : 0.61**	C : 0.33 SC : 0.50

If the summary is detected as facts contradict or express incorrect viewpoints in the first stage, then the score is ***0*** for the sufficient coverage score.

We first evaluate 18 articles using our two-stage scoring method, which is based on the four indicators described above. Table 3 presents the results of this evaluation.

Due to the evaluation bias in 'Correctness of ***VP***' and 'Comprehensible', we invited two researchers unfamiliar with *HotpotQA-topic* to participate in the scoring experiment for these two metrics. One of them is familiar with the ***NLP*** field but have no experience in the *HotpotQA-topic*, while one is a novice researcher unfamiliar with ***NLP***.

Table 3 shows the total results of the subjective evaluation. Our Few-shot method generally performs better in the most viewpoint-embedded summary. During the evaluation process, we made several notable discoveries.

1. The viewpoint 'limitation' in the paper is expressed subtly, making it difficult to identify. This results in all three methods performing less than satisfactorily. We also realized that the summary content for the 'performance' viewpoint is excessive. We need to further refine the structure of this viewpoint.

2. Although the Few-shot approach can get a brief and sufficient summary in most cases, its performance is mediocre in the viewpoint of 'dataset' and 'pre-training'. This is because the ***LLM*** mimics the format of Table 1 to achieve brevity, but it often overlooks crucial details and lacks a comprehensive understanding of the context. Conversely, the Zero-shot method tends to produce lengthy and less

effective summaries, as it lacks examples to guide the summarization process. However, in cases like 'Dataset' and 'Baseline', longer summaries may include more key information.

3. *Scispace* often generates summaries that use viewpoint-related vocabulary and their synonyms, but it does not always clearly convey the intended viewpoint-embedded information. This is similar to the issue of inadequate training in research. Furthermore, because there are no constraints on the input text, *Scispace* sometimes produces summaries from unrelated viewpoints. This issue can arise when extractive summarization is not performed. However, in the viewpoint - 'performance', this pattern actually enhances comprehensibility. From the viewpoint 'pre-training', we discovered that *Scispace* excels in mining paragraph chunking areas, capturing key information that predominantly using sentence chunks in this study may overlook. This is a direction we intend to improve in future research.

4. Examining the details of the subjective evaluation results presented in Table 6,7,8 reveals variations in the Comprehensible scoring among researchers, characterized by the following:

(1) All two researchers concluded that the summaries generated by *Scispace* contained more extraneous information, whereas our Zero-shot and Few-shot methods aligned better with the viewpoints. The Few-shot method, in particular, achieved a higher level of conciseness in the text.

(2) Researchers from fields unfamiliar with ***NLP*** may find the explanations of technical terms lacking in the Few-shot and Zero-shot methods, which can hinder their overall comprehension. In contrast, those with ***NLP*** experience have a foundation for analyzing these viewpoints. These concise summaries are particularly beneficial for them to conduct further survey.

(3) We also discovered that *Scispace*, lacking input text restrictions, generates content from previous issues in the viewpoint - 'limitation'. This is clearly erroneous, but novice researchers struggle to identify this error without reading the original paper.

5 Conclusion

This study proposes a ***diff-table*** system for Cross-sectional Research Insight Survey, aimed at aiding researchers in identifying similarities and differences in the research task through cross-comparison. Based on expert consensus, we consolidate and synthesize multiple papers with similar research objectives into a ***diff-table***. This table is created by **(1)** performing extractive summarization based on two-stage semantic text matching, and **(2)** generating abstractive summarization through two stages of prompt engineering. In the evaluation, we assessed the high consistency, correctness of viewpoint expression, comprehensible, minimal, and sufficient of the ***diff-table***, using objective measures such as *BERTScore* and subjective evaluations. Importantly, the ***diff-table*** holds potential for supporting Cross-sectional Insight Survey, providing a promising direction for future development. For future expansion and improvement of this study, the following points are proposed:

1. Machine learning technology for extractive summarization: This study used keyword scanning to extract sentences reflecting viewpoints. However, this method may struggle to identify sentences that do not align with our established rules, such as the sentence shown below that discusses previous issues that do not contain the keyword 'however'.

e.g. Previous issue: Since generators trained merely from recovering original statements are not encouraged to explore the possibilities of other reasonable statements.

To detect these irregularly expressed sentences, we need to create a viewpoint-based machine learning dataset for deeper viewpoint classification in the future. Furthermore, some key information, such as baseline of the pre-training model, is often found in the article's tables rather than in the body-text. Therefore, it is also important to identify and extract this kind of multi-modal information.

2. Expression of the structure of longitudinal knowledge: This study focuses mainly on the Cross-sectional Insight Survey. Based on these findings, the expression of the combination with the longitudinal knowledge structure is projected as an upcoming trend. Specifically, we will use the ***diff-table*** as a foundation and apply text similarity and citation relationships to establish connections between articles in the knowledge structure.

3. Enhance comprehensible for novice researchers: Enhance the narrative for novice researchers by fully explaining acronyms, offering concise descriptions of content, and including helpful annotations to aid their knowledge understanding. This requires a more refined prompt to produce diverse outputs that address the needs of novice researchers.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20H04295.

References

Roee Aharoni, Shashi Narayan, Joshua Maynez, Jonathan Herzig, Elizabeth Clark, and Mirella Lapata. 2023. Multilingual summarization with factual consistency evaluation. In Findings of the Association for Computational Linguistics: ACL 2023, pages 3562–3591, Toronto, Canada. Association for Computational Linguistics.

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2022. Automatic summarization of scientific articles: A survey. Journal of King Saud University-Computer and Information Sciences, 34(4):1011–1028.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers), pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Hainan Chen and Xiaowei Luo. 2019. An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. Advanced Engineering Informatics, 42:100959.

Po-Chun Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Categorizing citation relations in scientific papers based on the contributions of cited papers. In 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pages 384–389.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In Findings of the Association for Computational Linguistics: EACL 2023, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas D Cook, Donald Thomas Campbell, and William Shadish. 2002. Experimental and quasi-experimental designs for generalized causal inference, volume 1195. Houghton Mifflin Boston, MA.

Cheng Deng, Yuting Jia, Hui Xu, Chong Zhang, Jingyao Tang, Luoyi Fu, Weinan Zhang, Haisong Zhang, Xinbing Wang, and Chenghu Zhou. 2021. Gakg: A multimodal geoscience academic knowledge graph. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, CIKM '21, page 4445–4454, New York, NY, USA. Association for Computing Machinery.

İsmail Dönmez, Sahin Idin, and Salih Gülen. 2023. Conducting academic research with the ai interface chatgpt: Challenges and opportunities. Journal of STEAM Education, 6(2):101–118.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409.

Abdur Rahman Bin Md Faizullah, Ashok Urlana, and Rahul Mishra. 2024. Limgen: Probing the llms for generating suggestive limitations of research papers. arXiv preprint arXiv:2403.15529.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2023. What's new? summarizing contributions in scientific literature. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1019–1031, Dubrovnik, Croatia. Association for Computational Linguistics.

Naoya Inoue, Harsh Trivedi, Steven Sinha, Niranjan Balasubramanian, and Kentaro Inui. 2021. Summarize-then-answer: Generating concise explanations for multi-hop reading comprehension. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6064–6080, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jinghong Li, Wen Gu, Koichi Ota, and Shinobu Hasegawa. 2024a. Object recognition from scientific document based on compartment and text blocks refinement framework. 5(7).

Jinghong Li, Phan Huy, Wen Gu, Koichi Ota, and Shinobu Hasegawa. 2024b. Hierarchical tree-structured knowledge graph for academic insight survey. In 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pages 1–7.

Jinghong Li, Koichi Ota, Wen Gu, and Shinobu Hasegawa. 2023a. A text block refinement framework for text classification and object recognition from academic articles. In International

Conference on Innovations in Intelligent Systems and Applications, INISTA 2023, Hammamet, Tunisia, September 20-23, 2023, pages 1–6. IEEE.

JingHong Li, Huy Phan, Wen Gu, Koichi Ota, and Shinobu Hasegawa. 2024c. Fish-bone diagram of research issue: Gain a bird's-eye view on a specific research topic. arXiv preprint arXiv:2407.01553.

Jinghong Li, Hatsuhiko Tanabe, Koichi Ota, Wen Gu, and Shinobu Hasegawa. 2023b. Automatic summarization for academic articles using deep learning and reinforcement learning with viewpoints. The International FLAIRS Conference Proceedings, 36.

Meng-Huan Liu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Contributionsum: Generating disentangled contributions for scientific papers. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, pages 5351–5355.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4969–4983, Online. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. arXiv preprint arXiv:2303.15621.

Kaixin Ma, Hao Cheng, Yu Zhang, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. 2023. Chain-of-skills: A configurable model for open-domain question answering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1599–1618.

Kathleen McKeown and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D. Manning. 2019. Answering complex open-domain questions through iterative query generation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2590–2602, Hong Kong, China. Association for Computational Linguistics.

Md Mizanur Rahman, Harold Jan Terano, Md Nafizur Rahman, Aidin Salamzadeh, and Md Saidur Rahaman. 2023. Chatgpt and academic research: A review and recommendations based on practical examples. Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies, 3(1):1–12.

Giannis Vassiliou, Nikolaos Papadakis, and Haridimos Kondylakis. 2023. Summarygpt: Leveraging chatgpt for summarizing knowledge graphs. In European Semantic Web Conference, pages 164–168. Springer.

Juan David Velásquez-Henao, Carlos Jaime Franco-Cardona, and Lorena Cadavid-Higuita. 2023. Prompt engineering: a methodology for optimizing interactions with ai-language models in the field of engineering. Dyna, 90(230):9–17.

Xiaofeng Wang and Zhenshun Cheng. 2020. Cross-sectional studies: strengths, weaknesses, and recommendations. Chest, 158(1):S65–S71.

Jian Xu, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F Rousseau, Xin Li, Weijia Xu, Vetle I Torvik, et al. 2020. Building a pubmed knowledge graph. Scientific data, 7(1):205.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022. Prompt-based meta-learning for few-shot text classification. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1342–1357, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024. Cross-lingual cross-temporal summarization: Dataset, models, evaluation. Computational Linguistics, pages 1–44.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6197–6208, Online. Association for Computational Linguistics.

Table 4: Case study of ***diff-table*** for <Objective> : Sample paper (Qi et al., 2019)

Golden standard	- GOLDEN (Gold Entity) Retriever, it uses previous reasoning to generate a new query and retrieve evidence to answer the original question.
Our approach (Few-shot)	- Present GOLDEN (Gold Entity) Retriever. - Propose to rerank query results with a simple heuristic.
Our approach (Zero-shot)	- The paper introduces GOLDEN (Gold Entity) Retriever. - We propose to rerank query results with a simple heuristic to address the issue.
Scispace	- GOLDEN Retriever uses iterative reasoning for multi-hop question answering. - Queries generated for evidence retrieval enhance interpretability and scalability. - GOLDEN outperforms existing models on HOTPOTQA without BERT.

A Appendix

A.1 Table 4: Case study of diff-table

A.2 Table 5: Sample summary used in few shot prompt engineering

A.3 Table 6: Subject evaluation - Correctness of *VP* and Comprehensible

Evaluate multiple summary items in the cell one by one. Correctness measures how well the summary content matches the viewpoint. Comprehensibility measures the researcher's understanding of the overall content, measuring how well they can find in-depth survey clues.

1. **+2**: Mostly match = 80%-100%
2. **+1**: Medium match = 50%-80%
3. **-1**: Partially match = 20%-50%
4. **-2**: Does not match well = 0%-20%

Table 5: Sample summary used in few shot prompt engineering - Original text extracted from (Ma et al., 2023)

Viewpoint	Original text(Org) and its Sample Summary(S)	Feature
Previous issue	**Org** : However, this method suffers from undesirable task interference, i.e., negative transfer among retrieval skills. **S** : Suffers from undesirable task interference	Only emphasize the problem mentioned
Objective	**Org** : In this work, we propose Chain-of-Skills(COS), a modular retriever based on Transformer (Vaswani et al., 2017), where each module implements a reusable skill that can be used for different ODQA datasets. **S** : Chain-of-Skills(COS), a modular retriever based on Transformer.	Only extract fact author proposed
Dataset	**Org** : We consider six popular datasets for evaluation, all focused on Wikipedia, with four single-hop data, NQ (Kwiatkowski et al., 2019), WebQ, SQuAD and EntityQuestions **S** : Single-hop: NQ, WebQ, SQuAD, EntityQuestions.	Only extract the name of the dataset and its basic features
Pre-training	**Org** : For the second type, DPR-PAQ (Oguz et al., 2022) is initialized from the RoBERTa-large model (Liu et al., 2019b) with pretraining using synthetic queries (the PAQ corpus (Lewis et al., 2021) **S** : RoBERTa-large model with pretraining using synthetic queries	Only extract the name of the pre-training model and its basic features
Baseline	**Org** : For HotpotQA, we compare against three types of baselines, dense retrievers focused on expanded query retrieval MDR (Xiong et al., 2021b) and Baleen (Khattab et al., 2021)... **S** : Query retrieval MDR, Baleen, IRRR, TPRR	Only extract the name of the baseline and its basic features
Performance	**Org** : Our model, when coupled with the FiE, is able to outperform the previous baselines by large margins on OTT-QA, and we can see that the superior performance of our model is mainly due to COS. **S** : Outperforms previous baselines on OTT-QA, achieving superior performance due to COS.	Only extract the achievement author got
Limitation	**Org** : Our current COS's reranking expert only learns to rerank single-step results, thus it can not model the interaction between documents in case of multi-passage evidence chains. **S** : limited to reranking single-step results and cannot model interactions between documents in multi-passage evidence chains.	Only express something need to be improved
Future-work	**Org** :For future work, we are interested in exploring scaling up our method and other scenarios,e.g.,commonsense reasoning and biomedical retrieval. **S** : Scaling up , commonsense reasoning, biomedical retrieval	Only extract something will do in this future

Table 6: Subjective evaluation result of Few-shot - From 2 researchers, average score of random choosing 5 articles

	Researcher No.1 (Unfamiliar with *NLP)*		**Researcher No.2 (Familiar with *NLP)***	
	Correctness of *VP*	**Comprehensible**	**Correctness of *VP***	**Comprehensible**
Previous-issue	2	2	2	2
Objective	1.4	1.8	1	1.6
Dataset	0.4	1.4	0.75	0.75
Pre-training	2	2	1	0.33
Baseline	2	1.8	2	2
Performance	2	2	2	2
Limitation	1.5	2	2	2
Future-work	2	2	2	1.8
Average	1.66	1.88	1.59	1.56

Table 7: Subjective evaluation result of Zero-shot - From 2 researchers, average score of random choosing 5 articles

	Researcher No.1 (Unfamiliar with *NLP)*		**Researcher No.2 (Familiar with *NLP)***	
	Correctness of *VP*	**Comprehensible**	**Correctness of *VP***	**Comprehensible**
Previous-issue	2	1.8	2	1.6
Objective	1.4	1	1	1.6
Dataset	1.4	1.2	0.75	0
Pre-training	2	1.6	1	0.33
Baseline	1.4	1.6	1.6	1.8
Performance	2	1.8	2	1.8
Limitation	2	2	2	1.67
Future-work	2	2	2	1.8
Average	1.78	1.63	1.54	1.33

Table 8: Subjective evaluation result of Scispace - From 2 researchers, average score of random choosing 5 articles

	Researcher No.1 (Unfamiliar with *NLP*)		**Researcher No.2 (Familiar with *NLP*)**	
	Correctness of *VP*	**Comprehensible**	**Correctness of *VP***	**Comprehensible**
Previous-issue	0.8	1	0.4	0.4
Objective	2	1.6	2	1.4
Dataset	1.8	2	1.6	1.4
Pre-training	1	1	0.67	1.33
Baseline	1.6	0.6	1	0
Performance	2	2	2	2
Limitation	1.4	2	0	1
Future-work	2	1.6	2	2
Average	1.58	1.475	1.21	1.19