Doctoral Dissertation

A research survey assistant interface: An automatic generation of Top-down knowledge overviews for research topic from multiple academic papers

LI JINGHONG

Supervisor　Hasegawa Shinobu

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
Information Science

March 2025

Abstract

When starting new research, acquiring background knowledge is crucial to understand trends in the field. School lectures, seminars, and paper surveys are common ways to gain this knowledge. However, it can be challenging for novice researchers to apply knowledge from lectures to actual research. This difficulty arises when one does not habitually link concepts, assemble them logically, and analyze their meanings and reasons from different perspectives. Additionally, it's essential to grasp the forefront of cutting-edge science and technology. With the rapid progress in information science, self-study time has increased to understand the components and directions of research. Self-study can involve reading numerous papers, but it can be hard to extract important information from a vast amount of data for one's research. This process can be inefficient, and it's particularly challenging for novice researchers to identify crucial information due to lack of experience. Hence, providing novice researchers with a top-down grand-design — from a broad perspective to in-depth exploration to guide their research direction is a significant issue. To adopt top-down grand-design into research survey for novice researchers, this study is committed to establishing a research survey assistant interface, which ranges from a broader bird-eyes view to a deeper longitudinal & cross-sectional insight, to help novice researchers comprehend and explore research topics more logically. This system is divided into five main parts, ranging from data construction to interface development:

Phase 1 - Object Recognition from Scientific Document based on Compartment & Text Blocks Refinement Framework **(CTBR)** for infrastructure data generation: With the rapid development of the internet in the past decade, it has become increasingly important to extract valuable information from vast resources efficiently, which is crucial for establishing a comprehensive digital ecosystem, particularly in the context of research surveys and comprehension. The foundation of these tasks focuses on accurate extraction and deep mining of data from scientific documents, which are essential for building a robust data infrastructure. However, parsing raw data or extracting data from complex scientific documents have been ongoing challenges. Current data extraction methods for scientific documents typically use rule-based **(RB)** or machine learning **(ML)** approaches. However, using rule-based methods can incur high coding costs for articles with intricate typesetting. Conversely, relying solely on machine learning methods necessitates annotation work for complex content types within the scientific

document, which can be costly. Additionally, the data infrastructure in the subsequent phase development stage requires body text of high purity and low noise with clear divisions of the paper's internal elements. However, we've found that popular datasets like *S2orc* and *Unarxiv* sometimes intertwine images and tables within the body-text, causing discontinuity. This impurity could hinder the accuracy of subsequent semantic text analysis used for implementing research survey expression. From the perspective of analyzing the standard layout and typesetting used in the specified publication, we propose a new document layout analysis framework called **CTBR**. First, we conduct an in-depth exploration and classification based on the meanings of text blocks. Next, we utilize the results of text block classification to implement object recognition within scientific documents based on rule-based compartment segmentation. The object recognition results will be automatically incorporated to build the high-purity infrastructure data as the input of further processing.

Phase 2 - ***Fish-bone*** diagram of research issue: Gain a bird-eyes view on a specific research topic: Novice researchers often face difficulties in understanding a multitude of academic papers and grasping the fundamentals of a new research field. Existing keyword-based knowledge graphs make it difficult for researchers to deeply understand abstract concepts. Meanwhile, novice researchers may find it difficult to use *ChatGPT* effectively for research surveys due to their limited understanding of the research field. Without the ability to ask proficient questions that align with key concepts, obtaining desired and accurate answers from this large language model (**LLM**) could be inefficient. This subsystem aims to help novice researchers by providing a ***fish-bone*** diagram that includes causal relationships, offering an overview of the research topic. The diagram is constructed using the issue ontology from academic papers, and it offers a broad, highly generalized perspective of the research field, based on relevance and logical factors. This logical factor is established for novice researchers to navigate research topics, which expresses the route research topic $\rightarrow$ task $\rightarrow$ issue ontology $\rightarrow$ corresponding articles, serves as the foundation for in-depth data mining in a bird-eyes view survey. The issue ontology and these corresponding article nodes from ***fish-bone*** diagram will be imported as foundational information into the subsequent phases of research insight survey.

Phase 3 - Hierarchical Tree-structured Knowledge Graph for Academic Insight Survey: Depending solely on the ***fish-bone*** diagram in phase 2 makes it challenging for researchers to track changes and patterns over a specific period across multiple articles. Novice researchers, especially those without training in longitudinal perspective research—which tracks multiple relevant research branches over time and space. This lack of longitudinal insight

often hinders their ability to quickly grasp logical connections within relevant research tasks and discover new insights in a short time. One way to provide intuitive assistance to novice researchers is by offering relevant knowledge graphs (**KG**) and recommending related academic papers. However, existing navigation knowledge graphs mainly rely on keywords or meta information in the research field to guide researchers, which makes it difficult to clearly present the hierarchical relationships, such as inheritance and relevance between multiple related papers. Moreover, most recommendation systems for academic papers simply rely on high text similarity, confusing researchers as to why a particular article is recommended. They may lack the grasp of important information about the insight connection between 'Issue resolved' and 'Issue finding' that they hope to obtain. Hence, this subsystem aims to support research insight surveys for novice researchers by establishing a hierarchical tree-structured knowledge graph that reflects the inheritance insight and the relevance insight among multiple academic papers on specific research topics to address these issues.

Phase 4 - A Viewpoints Refinement ***diff-table*** System for Cross-sectional Insight Surveys In a Research Task : Relying solely on phase 3's longitudinal insight can make it difficult to identify similarities and differences among multiple articles. Therefore, we propose cross-sectional insight survey, aims to identify differences between groups, helping researchers understand various situations at certain time. This survey style outlines the fundamental attributes of the research task and the difference under these attributes. The advantage of this method is that the indicators are typically unified based on experts' consensus. However, the current knowledge graphs and automatic summarization systems used in research insight surveys seldom highlight the similarities and differences among multiple papers based on agreed-upon expert features. This can make it challenging for beginner researchers to understand the logical connections between research concepts. Therefore, this subsystem focuses on improving the extraction of differences among multiple articles related to the same task. This process expands the ***relevance tree*** knowledge through the conduct of Cross-sectional Insight Surveys. It offers a concise ***diff-table*** output format, tailored from the viewpoints of expert consensus. This subsystem aims to generating abstractive summarization based on the viewpoints of expert consensus and showing the differences under these consensuses. We created templates to embed these viewpoints in prompt description. These templates are used to generate an abstractive summarization for each cell in the ***diff-table***.

Phase 5 - Research Survey Supporting Interface: To enable researchers to understand survey logistics better and acquire more intuitive survey element prompts, we will incorporate the knowledge graph outputs from **Phases 2-4**

into a **UI** display, making it more engaging and easier to navigate through the vast array of information. Moreover, this interface will offer users the unique opportunity to explore and traverse through different survey paths to empower users to discover and choose the research direction that aligns best with their interests, objectives, and the scope of their work. Presenting the survey overview interface to novice researchers can leave a strong impression about the research topic, contributing to the creation of a 'learn how to learn' module. It's expected to become a significant research support tool for the upcoming generation of graduate students.

Keywords: Top-down, Research survey, Infrastructure data, Academic articles, Automatic summarization, Knowledge graph, Bird-eyes view, Longitudinal insight, Cross-sectional insight, Interface.

Acknowledgment

I would like to express my sincere gratitude to all those who have contributed to the completion of this dissertation. First and foremost, I thank Prof. Hasegawa, my supervisor of major research, for his invaluable guidance, support, and patience throughout this process. His expertise and insights have been instrumental in shaping this work.

I am deeply grateful for the generous financial support that enabled my attendance at various academic conferences. These opportunities not only broadened my horizons but also significantly fueled my enthusiasm for this research project.

I extend my heartfelt thanks to Prof. Inoue, my supervisor of minor research, for his guidance in developing the direction and conceptual framework for the diff-table component of this research.

I am also deeply grateful to Prof. Shirai for his helpful feedback and suggestions, which have significantly improved the quality of this dissertation.

Special thanks go to Prof. Gu for his assistance with the basic concepts and definitions in the research, such as issue ontology.

I would like to acknowledge the support of JAIST for providing the resources and facilities necessary for this research. Additionally, I thank my family and friends for their unwavering encouragement and moral support throughout this journey.

Finally, I extend my appreciation to all those who have directly or indirectly contributed to this work.

I sincerely hope that the research survey assistant interface will expand into more research branches across multiple directions, helping a wider range of researchers discover their ideal research paths.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Introductory remarks

A decade ago, the traditional research survey method was an extensive and time-consuming process. Researchers had to read numerous academic articles meticulously, mining their content to identify key research points and establish a comprehensive understanding of the field [74].

To build a foundation, researchers typically attended lectures to acquire background knowledge in their chosen domain [1]. Armed with this understanding, they would then use search engines to delve deeper, gathering in-depth knowledge [2–4]. This process allowed them to discern current trends and identify promising research directions that aligned with both their interests and the field's needs.

In recent years, academic research has undergone a significant transformation. The rapid development of digital libraries has created vast datasets, catalyzing the advancement of machine learning technologies [5]. These technologies now enable researchers to access professional knowledge more efficiently than ever before. Building upon these machine learning modules, powerful tools have been emerging that promise to revolutionize the research survey process. These include automatic summarization systems that distill key information from large volumes of text, paper recommendation engines that suggest relevant literature based on a researcher's interests, and sophisticated knowledge graph systems that visualize complex relationships between different concepts and studies [6–8].

The convergence of these technologies opens up exciting possibilities for conducting research surveys with unprecedented efficiency and depth. Researchers can now navigate vast bodies of literature more effectively, identify emerging trends more quickly, and gain insights that might have been overlooked using traditional methods [9]. This technological evolution isn't just changing how we conduct research surveys—it's expanding the boundaries of what's possible in academic inquiry.

However, tailoring the length of summaries, the scope of recommended

papers, and the volume of information in knowledge graphs to match users' needs remains challenging. This often results in a mismatch between the output and the user's specific requirements. When novice researchers enter a new field, they often struggle to navigate through large volumes of summarized data to find specific information they need. The challenge lies in striking a delicate balance: How to provide novice researchers with a comprehensive overview while also offering specific, in-depth investigation details and directions based on broader abstract concepts.

At this stage, novice researchers often need assist in organizing knowledge in a suitable survey format. Efficiently structuring information from extensive domain content is crucial work during survey process [10]. Domain knowledge typically progresses from broad, abstract concepts to specific, concrete examples. Capturing this hierarchical knowledge structure and presenting it visually in an intuitive manner poses a significant trend for next-generation research tools [11]. This hierarchical approach to knowledge organization enhances comprehension and aids in identifying research gaps and potential areas for innovation. By arranging information from general to specific, researchers can more effectively navigate complex topics, connect different direction of study, and uncover unexplored trend within their field [12]. Therefore, providing novice researchers with a visualization tool that supports top-down hierarchical surveys will enable them to emulate the expert's approach to learning. This will not only encourage novice researchers to conduct research surveys more efficiently but also stimulate their motivation to engage in the survey process.

1.2 Contrasting Survey Approaches: Novice Researchers vs. Research Experts

Master's and PhD students, who can be categorized as novice researchers in the field of computer science, often possess fragmented knowledge and practical experience with algorithms from their school course. They frequently encounter difficulties when initiating new research topics due to their lack of comprehensive research training and limited practical engineering experience. This combination of theoretical knowledge and practical gaps can create challenges in applying their learning to novel research contexts. During their research process, these novice researchers typically engage with various digital resources. They access digital libraries to read academic papers and essays or explore projects on platforms like GitHub and Hugging Face, and utilize search engines to find relevant information based on their existing

background knowledge [13]. Their goal is to discover research insights that align with their interests and expertise. However, the rapid advancement of information technology has led to an exponential growth in the number of research papers and projects available online. This information explosion makes it increasingly challenging for novice researchers to identify the specific information they seek within the vast digital libraries at their disposal [14].

The difficulty in finding desired information is further compounded by the novice researchers' limited project experience. They often struggle to break down their survey objectives into manageable tasks, which hinders their ability to approach the information they seek through a series of smaller, more achievable goals. In traditional academic settings, supervisors or experts would provide guidance by recommending specific papers and research directions to students. However, in the post-COVID-19 era, the availability of expert resources has become more limited [15]. The shift towards remote work and independent study has made self-directed learning an essential skill for these novice researchers to develop and refine. Even novice researchers with strong engineering skills may achieve quick interim results through their habitual bottom-up approach, tackling tasks one by one. However, only rely on bottom-up method [16] often leads to a lack of comprehensive understanding of the entire research field. By focusing excessively on engineering details and neglecting the overall research direction, they risk reinventing the wheel—their hard-earned achievements may have already been addressed by previous work. Consequently, even highly capable master's and doctoral students struggle to complete the entire research process independently. Therefore, supervisors need to invest time in guiding these students on how to conduct effective research surveys to help them find research originality [17]. One of the key strengths of experienced researchers is their ability to break down a broad research direction into multiple smaller, manageable objectives. This skill stems from their extensive domain knowledge and expertise. Even when venturing into an unfamiliar research field, these experts can leverage their experience gained from repeatedly managing diverse research projects. As a result, they have developed their own systematic approach to conducting research surveys. For instance, research experts possess the ability to identify the most opportune moments to gain a comprehensive bird eyes view of a research domain. They then establish a general framework of research branches, starting from broad, abstract concepts. From there, they delve deeper, exploring and constructing a detailed research route that reflects specific research issues. This methodical approach enables them to efficiently locate and analyze the most relevant papers for their study. This top-down approach to research surveys provides experienced researchers with a significant advantage. It enhances their ability to quickly decompose complex

tasks into more manageable components. Consequently, they can rapidly familiarize themselves with the key research points in new domains, allowing for more efficient and effective exploration of unfamiliar research territories. This skill is particularly valuable in today's fast-paced academic environment, where the ability to quickly adapt to new research domain is increasingly important.

1.3 After the boom of generative AI

After the COVID-19 pandemic, *ChatGPT* and its derivative paper reading assistance tools have gained widespread adoption [18]. These tools enable researchers to quickly locate and extract necessary information from literature through conversational interactions. Concurrently, the flourishing development of digital libraries has significantly enhanced the capabilities of generative Artificial Intelligence **(AI)**, drawing from large-scale paper datasets like *S2orc* [21]. These **AI** systems, powered by constantly updated, massive back-end datasets, are becoming increasingly sophisticated.

The advancements in generative **AI** for paper understanding have had far-reaching impacts. Primarily, they can produce high-quality summaries of single or multiple papers, dramatically improving researchers' efficiency in reading and comprehending academic papers [19]. For experts in a particular field, the flexible use of these tools can significantly amplify their productivity [20]. However, when it comes to supporting research surveys for novice researchers, existing **AI**-assisted tools still face several limitations:

1. As mentioned in **Section 1.2**, novice researchers often lack comprehensive domain knowledge, making it challenging for them to formulate suitable questions when using generative **AI** tools like *ChatGPT* [22,23]. Their limited familiarity with domain-specific terminology and insufficient research training hinder their ability to identify key research points across multiple relevant papers [24]. This difficulty manifests in several ways. Novice researchers struggle to grasp the overall landscape of multiple papers within a domain, making it harder for them to conduct top-down explorations of research branches, investigate longitudinal research directions, or analyze cross-sectional commonalities and differences. In contrast, expert researchers, having a solid grasp of these research elements and extensive training, can skillfully navigate these aspects. While it may be unrealistic to expect novice researchers to quickly become familiar with these advanced strategies like an expert, we can provide them with a template that mimics how research experts approach problems, thereby enhancing the educational value of existing

generative **AI**. This template could offer step-by-step guidance for conducting a top-down survey, providing relevant summary information at each stage to stimulate their curiosity and enhance their exploration process. Most existing generative tools focus on single functions such as reading comprehension or question answering. However, tools that offer top-down level assistance—mimicking the approach of expert researchers as described earlier—have yet to be developed.

2. Asking *ChatGPT* via chat-box to obtain domain knowledge and research trends can potentially lead to answers that are inconsistent with facts [25]. This is because current large language models like *ChatGPT* lack specific research training experience [24]. Consequently, the data retrieved from the underlying dataset may not fall within the scope of the question, resulting in generated answers that deviate from factual accuracy [26]. Such inaccurate and potentially hollow research guidance can negatively impact a novice researcher's survey process. Therefore, it becomes crucial to carefully control the range of text input into the Large Language Models **(LLM)** and set appropriate prompts to guide the **LLM**'s output effectively. Moreover, the limitations of these **AI** tools extend beyond just factual accuracy. They may struggle to provide the nuanced, context-specific insights that are often crucial in academic research. The **AI**'s responses might lack the depth and critical analysis that comes from years of hands-on research experience. This can be particularly problematic for novice researchers who may not yet have developed the skills to critically evaluate and contextualize the information they receive.
3. Existing generative **AI** tools primarily produce text-based summaries, lacking logical aids to enhance comprehension [27]. This limitation hinders novice researchers' ability to grasp the hierarchical structure of summarized information. When dealing with summaries of multiple articles, purely textual summaries can become unwieldy and fail to visually represent the interconnections and logical relationships between different papers. As a result, novice researchers often lose interest in engaging deeply with text-based summaries. In this case, a comprehensive route approach is needed that from broad, abstract concepts to detailed text summaries, inspiring novice researchers to analyze underlying reasons and gain deeper insights.

1.4 Reseach Objectives

With the emergence of *ChatGPT*, generative **AI** is revolutionizing the way early-stage researchers are educated. Based on this social background, considering the challenges novice researchers conducting extensive paper survey in new research fields, we formulate the following research question:

Main-RQ: How to effectively summarize large amount of articles within a research topic, from general concepts to specific details, to provide novice researchers with appropriate top-down research route?

→ **Objective:** This dissertation draws inspiration from the top-down research approach of senior research experts, aiming to develop a 'top-down' research survey assistant interface for novice researchers.

Specifically, this interface is characterized by a hierarchical structure that progresses from macro to micro: from ***fish-bone*** diagrams presenting broad abstract overviews, to more specific visualization forms such as ***relevance tree*** and ***diff-tables*** format. This visualization method can help novice researchers effectively capture key information in the research field within a short time, and connect these knowledge points to build their own knowledge system. Constructing such a comprehensive knowledge framework of the research field helps them to be more targeted when reading related papers subsequently, while also enabling them to ask more reasonable questions when using *ChatGPT*, thereby completing the early stages of the research survey effectively.

1.5 Dissertation outline

The structure of this dissertation is as follows:

Chapter 2 provides a comprehensive introduction to the research domains that have a significant influence on Top-down research surveys. This chapter delves into the various fields and sub-fields that contribute to the development and implementation of these survey methodologies.

Chapter 3 defines the requirements for the research survey assistant interface, presents the system design blueprint based on these definitions, and outlines the Front-end **UI** design and back-end data flow.

Chapter 4 provides a detailed introduction to the scheme for generating infrastructure data using the **CTBR** framework, forming the system's underlying data.

Chapter 5 introduces the first subsystem, the Fish-bone diagram for bird-eyes view. Built on the papers' issue ontology, this summary type is characterized by high integration, conciseness, and relative abstraction.

Chapter 6 details the second subsystem: the relevance-tree for longitudinal insight view. Building on the bird-eyes view, it delves deeper into abstract concepts at the individual paper level. This subsystem establishes a relevance-tree to illustrate connections between similar papers, effectively concretizing abstract summaries from a longitudinal perspective.

Chapter 7 presents the third subsystem, the Diff-table for cross-sectional insight view. Building on the relevance-tree, it further concretizes the longitudinal view cross-sectionally, exploring commonalities and differences within multiple papers.

Chapter 8 outlines the experimental design, methodology, and analysis for evaluating the research survey assistant interface. This subjective assessment specifically targets novice researchers, aiming to gauge the tool's effectiveness in supporting their research endeavors.

Chapter 9 summarizes the existing work in the field, analyzing the current state of research. It then discusses potential future directions and expansions that build on our achievements. This chapter consolidates our findings and offers insights into extending our work to address emerging challenges and opportunities in the domain.

Chapter 2

Literature review

In the field of Natural Language Processing (**NLP**), researchers often need to analyze large-scale text data for further development. This process requires efficient collection of substantial amounts of low-noise text. Once gathered, the data undergoes further processing—including extraction, classification, summarization, and visualization for practical applications. Existing work supporting research surveys from academic papers includes broad abstract knowledge graphs, content-tailored paper recommendation systems, information retrieval tools, and detailed paper summaries. Modern generative **AI**-based tools frequently combine several of these modules. **Section 2.2 - 2.6** explain the strengths and weaknesses of these existing works in research surveys to establish the direction of this dissertation.

2.1 Objective detection from academic article

Automatic collection of academic papers is essential for building a robust infrastructure for the survey content. Data mining technology is crucial for building the infrastructure data for research articles. With the advancements in text analysis tools for scientific documents, various techniques have been developed, such as extractive automatic summarization, abstract automatic summarization, and visualization slide generation using natural language processing [28]. In order to carry out the task mentioned above, data extraction techniques are essential to classify and identify information in the scientific document into various categories [29]. Document layout analysis **(DLA)** aims to detect and annotate the physical structure of documents [30]. However, parsing the structure and analyzing the content of scientific documents can be a complex and intricate task. Scientific documents often have irregular layout and the typesetting styles differ depending on the publication [31]. As a result, there is a tendency to extract fragmented data when attempting to retrieve internal information from scientific documents. For instance, the continuity of the main narrative within a file may be disrupted in mid-sentence caused from the reading order of individual text,

as well as interruptions such as figure titles, footnotes, and headers [32]. To address the aforementioned problem, rule-based conditional branching and regular expressions are commonly used to process text, involving parsing sentences and identifying keywords [31]. However, their ability to handle the complex structure used in scientific documents could be improved. This is because scientific documents often contain numerous patterns and irregularities that make them difficult to recognize by traditional algorithms. In order to tackle this problem, researchers have explored the machine learning algorithms, which are more flexible and able to learn based on the presented data. By adopting machine learning methods, it becomes possible to take into account the specific patterns in scientific documents, thereby enhancing the accuracy and effectiveness of automatic text processing [33]. However, scientific documents may have varying fonts and typesetting, which require more complex vectorization and annotation methods to increase generality. Formulating these intricate techniques can be a time-consuming and costly endeavor [78]. **Sections 2.1.1 - 2.1.3** will provide a summary of representative research in **DLA**, highlighting their strengths and limitations.

2.1.1 Existing research on layout analysis of scientific document *PDFs*

Cheng et al. present a two-stream multi-modal Vision Grid Transformer for document layout analysis, which directly models 2D token-level and segment-level semantic understanding [34]. Lopez et al. developed GROBID, which extracts the bibliographical data corresponding to the header information (title, authors, abstract, etc.) and to each reference (title, authors, journal title, issue, number, etc.) [35].

2.1.2 Recognition of figures and tables in scientific article *PDFs*

Clark et al. developed PDFFigures 2.0, which extracts figures, tables, and captions from computer science articles in *PDF* format. The algorithm analyzes the structure of individual pages by detecting captions, graphical elements, and chunks of body-text and then locates figures and tables by reasoning about the empty regions within that text [36]. Frerebeau et al. provided Tabula, a web-based system that extracts tables from untagged *PDF* documents. It uses customizable heuristics to detect tables and reconstruct cell structure based on text and ruling lines in the *PDF* [37].

2.1.3 Recognition of figures and tables in images of various formats

Smock et al. created PubTables-1M, a new dataset for table extraction from scientific articles. It has almost one million tables, detailed headers, and location information for table structures. PubTables-1M also supports multiple input modalities and is useful for many modeling approaches. Their transformer-based object detection models trained on PubTables-1M produce excellent results for detection, structure recognition, and functional analysis without any special customization [38]. Paliwal et.al built a novel end-to-end deep learning model for both table detection and structure recognition. The model exploits the interdependence between the table detection and table structure recognition to segment the table and column regions [39].

2.1.4 Position of CTBR for infrastructure data building

The **Table 2.1** presents a summary of the previous research and their limitations on data extraction for scientific documents. Previous studies have faced difficulty in accurately distinguishing text blocks within figures and tables. This study aims to address these limitations by refining text blocks using different features, including position, size, line and column spacing, font type, and font size. By considering these features and compartment segmentation, we can enhance the ability to distinguish various objects within scientific documents. The precise identification of figures, tables, and body text using **CTBR** enables us to create a more refined local dataset of research papers

2.2 Research survey through academic knowledge graph

Novice researchers need to comprehensively understand the knowledge overview in their field to break down research tasks through research surveys. The development of knowledge (***KG***) can benefit for novice researchers in efficiently comprehending and exploring research topics. As previous studies on the ***KG*** generated from academic papers, Xu et al. constructed a PubMed knowledge graph that includes meta information such as bio-entities and authors [40] and Martha et al. proposed the Tree of Science ***(ToS)***, which recommends articles based on their position in the graph of citation [41]. However, for novice researchers, relying solely on meta or citation information is insufficient to provide enough information about the research content in

Table 2.1: The Summary of previous work on data extraction for scientific documents

Reference	**Method**	**Feature/Advantage**	**Limitation**
VGT (2023) [34]	Grid Transformer	• New diverse and detailed manually-annotated dataset D^4LA • In-depth layout analysis • Pre-trained for 2D token-level and segment-level semantic understanding	If there is a cluster of text blocks in a figure/table, it can be difficult to distinguish whether it is body text or figure/table region.
Table Transformer (2022) [38]	Detection Transformer	Large training data, and supports multiple input modalities and is useful for many modeling approaches.	Miss detection in complex pattern matching involving continuous tables.
Table net (2020) [39]	End-to-end Deep learning	This model uses interdependence between table detection and structure recognition tasks to segment table and column regions.	Header of the table is difficult to fit.
Pdffigure 2.0 (2016) [36]	Rule-based	Analyzes page structure and locates figures and tables by analyzing empty regions within text.	Miss detection occurred when figures and tables appear continuously.
Tabula (2018) [37]	Rule-based	It uses customizable heuristics to detect tables and reconstruct cell structure based on text and ruling lines in the *PDF*.	If recognition is not restricted to the table zone, the body-text and section title patterns may be mixed with the detected result.
Grobid (2008-2023) [35]	CRF	Extract and reorganize not only the content but also the layout and text styling information.	It is difficult to extract noise-free text because information with figures, tables, and equations are included in the body-text.

that field to gain the overview concept map of the research branch. Deeper data mining from the academic paper is still required to refine the academic knowledge graph, which may provide high-quality guidance for the research survey. The knowledge graphs that involve internal information of academic papers include Chan et al. proposed representation ontology for a four-space integrated ***KG*** (background, objectives, solutions, and findings) using NLP technology, as well as Tu et al. proposed Semantic Knowledge Graph ***(SKG)*** that integrates semantic concepts to represent the corpus [42] [43]. They both classify the content of papers into entity concepts and associated research key-points in their knowledge graphs. However, the association paths of complex concept maps can be cumbersome, making it difficult for novice researchers to accurately locate academic papers in a specific research direction, which may result in limited insight into the research branch. Thus, extracting insightful content from the paper is crucial to guide researchers in conducting effective surveys. Moreover, integrating cited material from the contribution text can enhance the summarization by covering multiple papers, such as Chen et al. propose the task of citation relation classification based on the contributions of cited papers to improve the summarization system for scientific papers [44]. These works commonly summarize the part of contribution and related text that provide a road-map for the research survey. However, focusing solely on the contribution may ignore the source of research motivation — what gaps in previous studies prompted the author's contribution? Furthermore, covering the entire research topic comprehensively through the part of contribution alone is insufficient, which makes it difficult to understand the research overview. Conversely, previous research that generates a knowledge graph may better support understanding the research overview, such as Chan et al. proposed a representation ontology for an integrated four-space keyword based Knowledge Graph (background, objectives, solutions, and findings) using Natural Language Processing technology. However, the 'abstract' of the paper provides insufficient information to understand the detailed research issue fully. For instance, the part of 'abstract' seldom underscores the element of improvement from previous research, making it difficult to reflect the clue of author's motivation. Duan et al. generated summarization of the development of the research field over nearly a decade based on keywords, which played a key role for novice researchers to grasp the direction of the research field [45]. However, the keyword-based knowledge graph might be too vague for novice researchers to understand the inherent logical connections among multiple academic papers. Zhang et al. present a heterogeneous network literature recommendation method based on the domain knowledge graph and hotspot information composition [46]. However, relying solely on hotspot information is insufficient to delve into the historical issues remaining in the

research domain and whether these issues have been resolved or not.

2.3 Research survey through paper recommendation

As online information grows exponentially, recommender systems play a crucial role in mitigating information overload [47]. Similarly, novice researchers need to efficiently select relevant papers from the vast pool of online information in their field through comprehensive research surveys. Thus, Recommender systems have become a key tool for personalized content filtering in education [48]. In academic world, inspiring new directions and understanding state-of-the-art approaches are important aspects of research survey assistance [49]. Following this, paper recommendation systems are now widely used to identify specific papers within a given field based on particular requirements [51]. Pinedo et al. introduced ArZiGo, a Web-based scientific paper recommendation system. It utilizes the Semantic Scholar Open Research Corpus, a growing multidisciplinary literature database, to provide users with personalized paper suggestions [50]. This approach provides personalized paper recommendations based on user characteristics, which requires users to share some personal information. However, two shortcomings of this method: It requires users to disclose personal data and does not offer a comprehensive overview that would help users identify their preferred research directions. Chaudhuri et al. identifies four key features for representing research articles: keyword diversity, text complexity, citation trends, and scientific quality. These features aim to enhance recommendation variety, match papers to readers' comprehension levels, assess relevance over time, and evaluate overall quality [52]. However, the absence of a similarity metric between recommended papers prevents novice researchers from understanding internal connections among these papers. Such connections might represent expert consensus in the field or highlight similar research issues across multiple studies. Consequently, it is crucial to present users with a global perspective that highlights both the relevance among the selected papers in a fine-grained conditions, thereby inspiring their exploratory instincts.

2.4 Research survey through information retrieval

On the other hand, Information Retrieval ***(IR)*** plays a crucial role in helping users explore the information they consider important, similar to the internet browsing and question-answer systems [53]. Musaev et al. improve the efficiency of searching for scientific and educational resources using **AI** techniques within the intelligent information system *SMART TUIT* [54]. However, the *SMART TUIT* subsystem's synergistic effects fail to connect clues based on the characteristics of a specific domain's paper. Moreover, its information processing steps are excessively complex, potentially causing retrieved information to stray from the research field's main focus. Sharma et al. presented a novel hybrid semantic indexing approach for unstructured text documents, combining machine learning techniques with domain ontology [55]. However, relying solely on domain ontology fails to adequately capture the intricate connections between deep-layer issue ontology in academic papers. This limitation hinders researchers from effectively identifying commonalities and differences among papers through research issue in a given field, making it difficult to discern distinctions differences in multiple papers with similar research objectives. Consequently, it becomes challenging to quickly uncover new research findings based on these differences.

2.5 Research survey through automatic summarization

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 [56–58] or excavate the inheritance relationship of the paper itself [59]. 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 [40, 46, 60], 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 [61–64]. 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.

2.6 Exist AI tools for research survey

The advent of large language models (**LLM**) like *Chatgpt* have opened up new possibilities for utilizing these extensive data sources in **NLP** tasks, including question-answering systems and automatic summarization for academic papers. This advancement enables researchers to conduct research surveys more efficiently using research tools and resources, such as *ChatPDF* and *SciSpace*, which have been developed based on *ChatGPT* and specifically tailored for academic papers understanding [65]. Among these:

- **QA with paper**: Several existing tools focus on understanding individual papers. For example, *ChatPDF*[1] [66] and Chat-Paper[2] offer user interfaces that enable researchers to interact directly with academic papers for concise summarization. This approach particularly benefits researchers who might struggle to efficiently parse dense academic texts. Skarlinski et al. developed PaperQA2, an **AI** system that outperforms human experts in scientific research tasks. The system excels in answering questions from single scientific literature, creating more accurate summaries than Wikipedia, and identifying contradictions across scientific papers at scale [67, 68]. PaperQA2 has nearly perfected summarization for individual papers, producing summaries that closely mirror human thought patterns. However, to obtain refined summaries, novice researchers must engage in iterative question-answering interactions with PaperQA2. This process is difficult for them because they lack the upstream knowledge frameworks necessary to formulate effective questions quickly.
- **Question-answering literature search tools**: *SciSpace*[3] and *Elicit*[4] [69] enables literature searches using query-related keywords. Results

[1] https://www.chatpdf.com
[2] https://chatpaper.com
[3] https://typeset.io
[4] https://elicit.com

can be sorted by title, abstract, PDF availability, year, and citations. It also offers tabular displays of research findings and methods, facilitating cross-sectional paper comparisons using **AI**-powered literature survey tools. *Consensus*[5] is a tool that aids initial literature searches in new fields. Meta-analyses and systematic reviews provide concise research overviews, offering a quick grasp of the field. These sources not only present relevant content from multiple papers but also synthesize findings into comprehensive summaries, enabling researchers to efficiently understand the research insight. However, **LLMs** often lack specialized research training and refined prompts tailored to academic contexts. Moreover, the input text range is not optimally adapted for specific research perspectives. Consequently, **LLM** outputs frequently include extraneous or irrelevant information that does not align with the desired research focus. This lead to overly verbose summaries that are challenging for researchers to efficiently process and identify key-points.

- **Research mapping tool**: Research mapping tool visualizes citation networks, referenced works, and similar publications as a network diagram of nodes and edges, facilitating in-depth longitudinal exploration across multiple research papers. *ResearchRabbit*[6] [70] and *Connected Papers*[7],[8] excels in visualizing connections between research papers and authors. Its network-style graph illustrates citation relationships and co-authorship patterns, enabling users to easily trace research developments and identify influential works. *Litmaps*[9] [71] is a mapping tool that uses 'Discover to find related literature from multiple sources. It suggests papers with citation and reference relationships, as well as related works, and allows users to check abstracts within the Litmaps interface. However, these tools often lack sophisticated metrics for calculating similarity between papers and fail to provide visualizations based on such metrics. Furthermore, they don't offer explanations for the connections between nodes. Without clear descriptions of paper relationships and concise summaries, novice researchers may struggle to grasp the broader context and evolution of research topics. This limitation hinders researchers' ability to develop a comprehensive understanding of the subject. Consequently, novice researchers find it challenging to conduct in-depth longitudinal explorations and create

[5] https://consensus.app
[6] https://www.researchrabbit.ai/team
[7] https://github.com/ConnectedPapers
[8] https://www.connectedpapers.com
[9] https://www.litmaps.com

their own comprehensive survey conceptual maps. Moreover, these tools still require users to employ advanced keywords to retrieve relevant articles. Novice researchers, lacking domain-specific knowledge, often struggle to use appropriate technical terms when formulating questions. This type of uninformed search, without proper guidance, leads to inefficient longitudinal explorations of a research topic.

Table 2.2 compares the research survey-related functionalities between existing AI research tools and our Research Survey Assistant Interface.

2.7 Position of our research survey assistant interface:

Unlike previous research with its limited and isolated functional systems, our research survey assistant interface aims to establish a comprehensive and continuous support system. Following a top-down approach, it progresses from abstract concepts to in-depth summaries of specific papers. This design plays a crucial role in supporting novice researchers during the early stages of their work, offering a seamless transition from broad overviews to detailed analyses. **Table 2.3** summarizes the limitations of existing research and tools for supporting research surveys, and we summarizes the research questions and solutions this study proposes to address these shortcomings:

1. How to provide novice researchers with a hierarchical bird's-eye view that reflects the causal-effect in a research topic?
   → We develop a new type of the ***fish-bone*** diagram, generated by internal issue ontology of academic papers function on reflecting logical guide chain of topic → task → issue ontology is established for novice researchers to navigate research topics.

2. How to help researchers gain effective insights by identifying connections between the evolution and relevance across multiple research papers?
   → We create a tree-structured knowledge graph for Academic insights survey from multiple academic papers on a specific research topic, function on expanding the citation inheritance and relevance associations based on academic issue ontology.

3. How to aid novice researchers in identifying similarities and differences in the research task through cross-comparison?
   → We develop a system that assists researchers in the cross-sectional

Table 2.2: Functions list - Comparison of AI Research Tools via research survey perspectives

	Paper recommendation	Summarization of single paper	Table-based summarization	Paper relevance analysis	Logical conceptual survey route	Citation mapping	Keyword-based knowledge linking
Paper Qa2	×	◯	◯	×	×	×	×
Scispace	◯	◯	◯	×	×	×	×
Elicit	◯	◯	◯	×	◯	◯	×
Chatpdf	×	◯	×	×	×	×	×
Consensus	◯	◯	◯	×	◯	◯	×
Research Rabbit	◯	×	×	◯	×	◯	×
Connect papers	◯	×	×	×	×	◯	×
Litmaps	◯	×	×	◯	×	◯	×
OpenScholar [a]	◯	×	×	×	◯	×	×
Research survey assistant interface [b]	◯	◯	◯	◯	◯	◯ [c]	◯

[a] https://openscholar.allen.ai

[b] While our research survey assistant interface includes all the functionalities mentioned above, it currently lacks the chatbot-based question-answering capabilities found in these **AI** tools.

[c] While we implemented citation mapping in the inheritance tree component of the insight tree in **Chapter 6**, we used a simplified citation information representation instead of the inheritance tree for the experiments in **Chapter 8**.

research insight survey through abstractive summarization in a viewpoints-embedded ***diff-table*** format.

4. How to effectively summarize large amount of articles within a research topic, from general concepts to specific details, to provide novice researchers with appropriate top-down research route?
   → We develop a top-down research survey assistant interface for novice researchers. Our system provides layered, logically structured survey support, ranging from broad, abstract ***fish-bone*** summaries to in-depth, detailed diff-table summaries, providing a hierarchical, comprehensive overview among large amount of papers.

Table 2.3: The Summary of Previous Work of research survey method

Method	**Role in Top-down Survey**	**Feature/Advantage**	**Limitation**
Academic KG	• Bird's-eye view • Longitudinal insight view	• 1. Meta Information analysis • 2. Citation Relations • 3. Network building by certain keyword • 4. Hotspot research trend exploration	• 1. Insufficient comprehensive overview of research content • 2. Reasons for citations are not reflected • 3. Complex concept maps with intricate association • 4. Fails to reflect the origins of hotspot research and its associated research issues
Paper recommendation	• Longitudinal insight view	• 1. Provide users with personalized paper suggestions • 2. Keyword diversity, text complexity, citation trends, and scientific quality	• 1. Fails to highlights the relevance among the selected papers • 2. Absence of a similarity metric between recommended papers
Information Retrieval	• Longitudinal insight view	• 1. Helping users trace the information • 2. Semantic indexing approach with domain ontology	• 1. Fail to connect clues of a specific domain's paper • 2. Fails to capture the intricate connections between deep-layer issue
Automatic summarization	• Longitudinal insight view • Cross-sectional Insight view	• 1. Trace research progression • 2. Provides viewpoints-insight into the research direction	• 1. Difficult to compare differences in details • 2. The expression of the limited viewpoints
Generative AI tools	• Longitudinal insight view • Cross-sectional Insight view	• 1. Produces summaries that similar to human experts • 2. Findings into comprehensive summaries among multiple papers. • 3. Network-style graph illustrates citation relationships.	• 1. Novice researchers are hard to quickly formulate effective questions. • 2. Outputs frequently include irrelevant information • 3. Lack sophisticated metrics for calculating similarity

Chapter 3

Methodology

3.1 Definition in Top-down Survey

A top-down research survey is a methodical approach to conducting research investigations through multiple papers in a specific research field. It begins with a broad overview of a topic and gradually narrows down to specific details as a survey route. This approach forms the foundation of our survey assistant interface. Here, the top-down survey process is shown in **Figure 3.1**. In this section, we introduce the definition and key components of the top-down research survey method.

3.1.1 Research topic and Research task

A research topic is derived from keywords related to a specific research field and serves as a focal point for academic inquiry within a particular domain of study. It often emerges from the intersection of multiple concepts or areas of interest, providing rich ground for investigation and analysis. Multiple research tasks emerge within a research topic, each representing a distinct aspect of the broader subject. As researchers delve into these tasks, they uncover diverse research directions—pathways of inquiry that lead to new discoveries, theories, or applications in the field.

3.1.2 Issue ontology

Ontology represents aspects of the world through defined classes of 'entities' and their relationships. Used in various fields, they provide clarity, consistency, and data linkage. Ontology shows promise for revolutionizing study descriptions, findings, and theory expression [72,73]. To link the relationships between academic issues across multiple academic papers, we defined an ontology with academic issues as its foundation, establishing data linkage in the research topic. The specific definition is as follows:

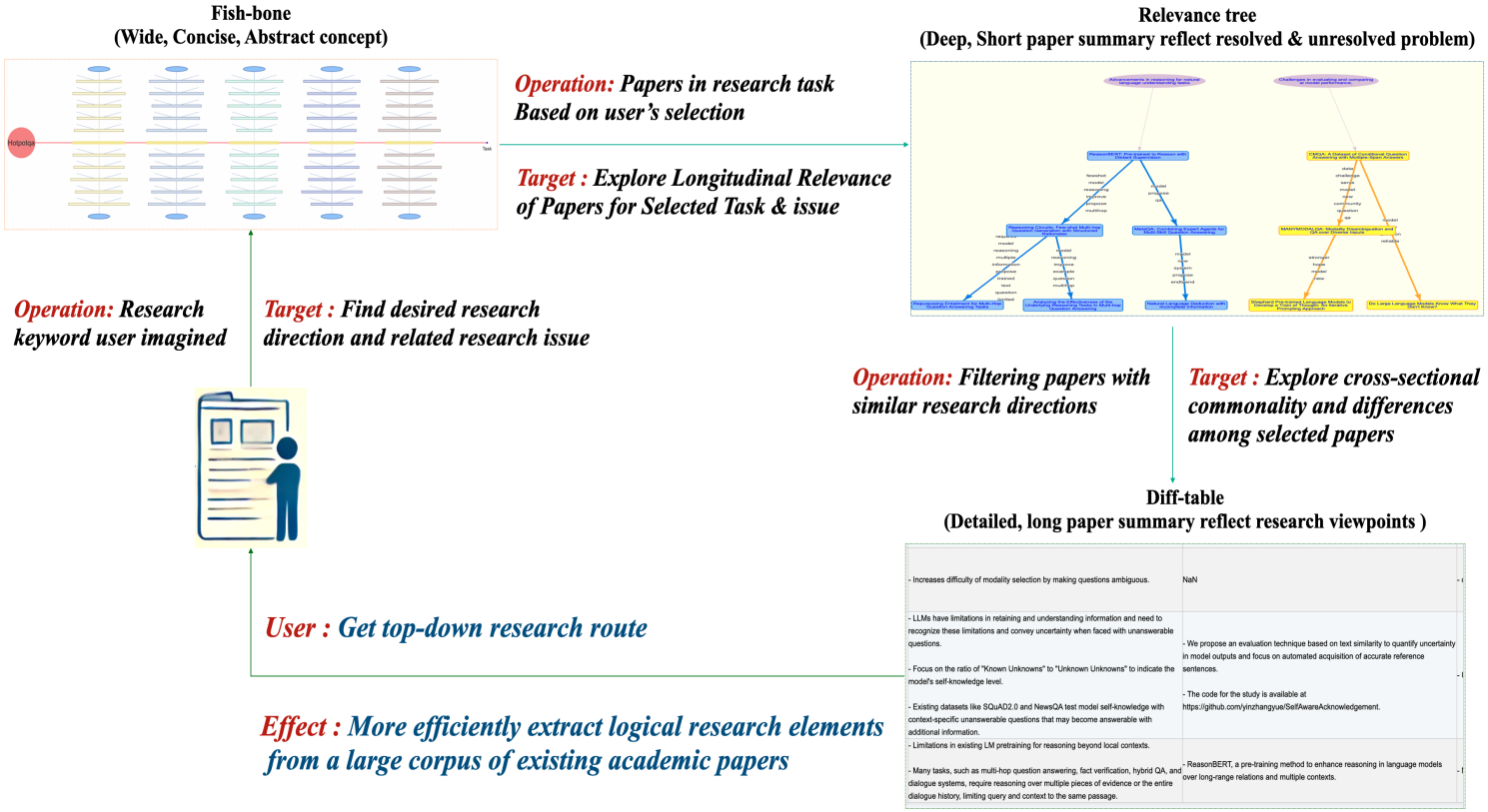


Figure 3.1: Top-down survey process

In academic papers, certain key elements provide insight into the author's reasoning and thought process. These elements reveal how the author discovers problems, investigates them, and contemplates solutions. This can be regarded as the writing clues of the article. The related sentences that appear throughout the papers make up the clues, which are crucial for readers to grasp the bird-eyes view of the research topic. These critical sentences encompass the concept of 'issue ontology'. Issues embody various debates within the academic world. Some of these have been resolved, some have been identified but remain unresolved, and others still require optimization and improvement.

By analyzing the relationships and logical connections between issue ontology across multiple papers, we can offer valuable assistance to novice researchers. This approach helps them construct a more systematic and structured knowledge base within specific research topics while also guiding them towards potential research directions. Understanding the connections between issues in various literature enables researchers to gain a clearer overview of the research field, identify key problems and research gaps, and receive crucial guidance for positioning their own research and selecting appropriate directions.

3.1.3 Viewpoint

Viewpoints in research refer to the established and widely accepted research methodologies and perspectives within a specific academic domain. These viewpoints represent a collective understanding that has evolved over time, from the initial stages of the research area to its current state, culminating in a cohesive and unified perspective [74]. This consensus among experts forms the foundation upon which most scholarly works in the field are built, providing a common framework for analysis and discussion.

The majority of academic papers within a given research domain are structured around these specific viewpoints, reflecting the shared understanding and methodological approaches of the field. These viewpoints serve as guiding principles, shaping the way research questions are formulated, data is collected and analyzed, and conclusions are drawn. They provide a lens through which researchers examine and interpret phenomena within their domain of study.

3.2 System requirement : Top-down view of survey

Regarding the research questions we mentioned in **Chapter 1** : How to effectively summarize large amount of articles within a research topic, from general concepts to specific details, to provide novice researchers with appropriate top-down research pathways?

→We develop a top-down research survey assistant interface for novice researchers. Our system provides layered, logically structured survey support, ranging from broad, abstract ***fish-bone*** summaries to in-depth, detailed diff-table summaries, providing a hierarchical, comprehensive overview among large amount of papers. The research survey assistant interface is divided into four main parts to form the System requirement

3.2.1 Bird-eyes view survey

RQ-sub1: How to provide novice researchers with a hierarchical bird-eyes view that reflects the causal-effect overview in a research topic?

→ We develop a new type of the ***fish-bone*** diagram, generated by internal issue ontology of academic papers function on reflecting logical guide chain of topic → task → issue ontology is established for novice researchers to navigate research topics.

Just as birds overlook a city's intricate layout from above, taking in skyscrapers, open spaces, mountains, and lakes, a similar panoramic view benefits novice researchers starting their academic exploration. By providing a comprehensive, high-level view of the research landscape that help them easily grasp the logic and connections within research topic. This broad overview illuminates the context of their research topic and reveals the driving forces behind various research directions in that domain.

This methodology is positioned at the top level of the research survey assistant interface, offering a comprehensive overview of the research topics and their associated directions.

3.2.2 Longitudinal insight view survey

RQ-sub2: How to help researchers gain effective insights by identifying connections between the evolution and relevance across multiple research

papers?

→ We create a tree-structured knowledge graph for Academic longitudinal insights survey from multiple academic papers on a specific research topic, function on expanding the citation inheritance and relevance associations based on academic issue ontology.

Longitudinal survey is a research approach that allows for in-depth cause-and-effect comparisons across multiple papers. This method enables researchers to meticulously track changes, identify emerging patterns, and observe evolving trends within a specific research task over an extended period. By examining a series of studies conducted at different points in time, researchers can gain valuable insights into the progression of ideas, methodologies, and findings within their direction of interest. This approach is particularly useful for understanding how research questions, hypotheses, and providing a comprehensive view of the research task's evolution. For example, they track developments from ancient times to now, identify shifts in user interests and capture their changes through time [75] or excavate the inheritance and relevance relationship of the paper itself [59]. This methodology is positioned at the medium level of the research survey assistant interface, situated beneath the bird-eyes view survey. It offers a comprehensive insight into research tasks, revealing their relevance and evolution over time.

3.2.3 Cross-sectional insight view

RQ-sub3: How to aid novice researchers in identifying similarities and differences in the research task through cross-comparison?

→ We develop a system that assists researchers in the cross-sectional research insight survey through abstractive summarization in a viewpoints-embedded ***diff-table*** format.

The Longitudinal insight view tracks the evolution of research paths over time, focusing on paper-based longitudinal expansion that follows the temporal development of research tasks. In contrast, the Cross-sectional view facilitates intuitive comparisons of multiple papers within a research task, using common criteria specific to the research domain. Cross-sectional study aims to identify differences between groups, helping researchers understand various situations at certain time [76]. 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 based on 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. This methodology occupies the bottom level of the research survey assistant interface, positioned beneath the longitudinal insight view survey. It provides a comprehensive analysis of multiple papers, highlighting their commonalities and differences.

3.3 System Overview

Our designed interface encompasses the entire process from fundamental infrastructure data construction to visual summarization presentation. Based on the background presented in **Chapter 1 - 2** and the system requirements outlined in **Section 3.2**, we divide the Research Survey Assistant interface into four distinct components. The following outline provides a clear overview of each integral element contributing to our system's functionality and effectiveness:

3.3.1 Infrastructure data building

With the rapid development of the internet in the past decade, it has become increasingly important to extract valuable information from vast resources efficiently, which is crucial for establishing a comprehensive digital ecosystem, particularly in the context of research surveys and comprehension. The foundation of these tasks focuses on accurate extraction and deep mining of data from scientific documents, which are essential for building a robust data infrastructure. However, parsing raw data or extracting data from complex scientific documents have been ongoing challenges. Current data extraction methods for scientific documents typically use rule-based **(RB)** or machine learning **(ML)** approaches. However, using rule-based methods can incur high coding costs for articles with intricate typesetting. Conversely, relying solely on machine learning methods necessitates annotation work for complex content types within the scientific document, which can be costly. Additionally, few studies have thoroughly defined and explored the hierarchical layout within scientific documents. The lack of a comprehensive definition of the internal structure and elements of the documents indirectly impacts the accuracy of text classification and object recognition tasks. From

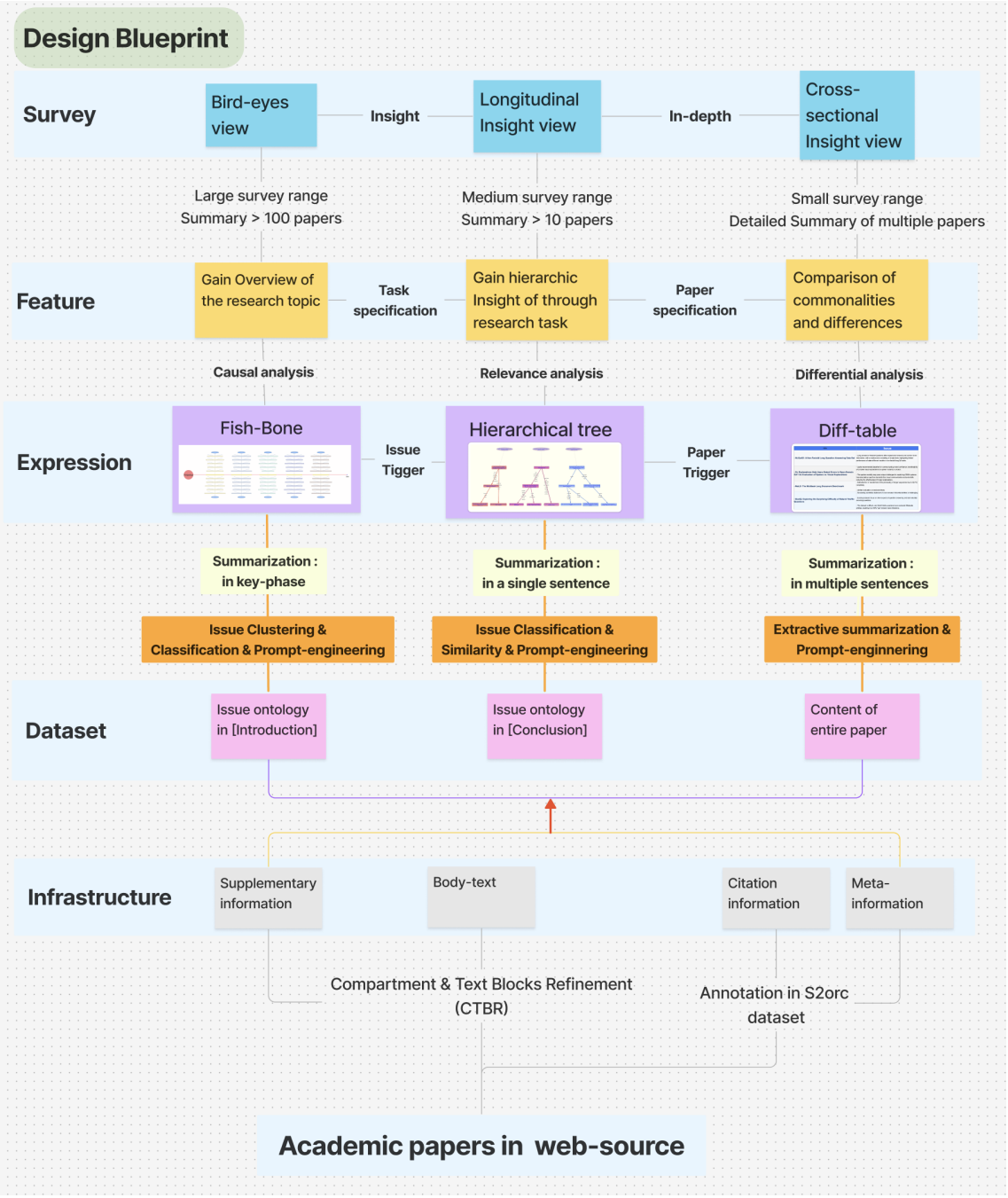


Figure 3.2: Design Blueprint

the perspective of analyzing the standard layout and typesetting used in the specified publication, we propose a new document layout analysis framework called Compartment & Text Blocks Refinement **(CTBR)**. Firstly, we define scientific documents into hierarchical divisions: base domain, compartment, and text blocks. Next, we conduct an in-depth exploration and classification of the meanings of text blocks. Finally, we utilize the results of text block classification to implement object recognition within scientific documents based on rule-based compartment segmentation.

3.3.2 Subsystem 1 - Fish-bone for bird-eyes view

Novice researchers often face difficulties in understanding a multitude of academic papers and grasping the fundamentals of a new research field. To solve such problems, the knowledge graph supporting research survey is gradually being developed. Existing keyword-based knowledge graphs make it difficult for researchers to deeply understand abstract concepts. Meanwhile, novice researchers may find it difficult to use *ChatGPT* effectively for research surveys due to their limited understanding of the research field. Without the ability to ask proficient questions that align with key concepts, obtaining desired and accurate answers from this large language model *(***LLM***)* could be inefficient. This study aims to help novice researchers by providing a ***fish-bone*** diagram that includes causal relationships, offering an overview of the research topic. The diagram is constructed using the issue ontology from academic papers, and it offers a broad, highly generalized perspective of the research field, based on relevance and logical factors. Furthermore, we evaluate the strengths and improvable points of the ***fish-bone*** diagram derived from this study's development pattern, emphasizing its potential as a viable tool for supporting research survey.

3.3.3 Subsystem 2 - Relevance for longitudinal insight view

Research surveys have always posed a challenge for novice researchers who lack research training. These researchers struggle to understand the directions within their research topic and the discovery of new research findings within a short time. One way to provide intuitive assistance to novice researchers is by offering relevant knowledge graphs (***KG***) and recommending related academic papers. However, existing navigation knowledge graphs mainly rely on keywords or meta information in the research field to guide researchers, which makes it difficult to clearly present the hierarchical relationships, such as inheritance and relevance between multiple related papers.

Moreover, most recommendation systems for academic papers simply rely on high text similarity, confusing researchers as to why a particular article is recommended. They may lack the grasp of important information about the insight connection between 'Issue resolved' and 'Issue finding' that they hope to obtain. This study aims to support research insight surveys for novice researchers by establishing a hierarchical tree-structured knowledge graph that reflects the inheritance insight and the relevance insight among multiple academic papers on specific research topics to address these issues.

3.3.4 Subsystem 3 - *Diff-table* for cross-sectional insight view

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 generating 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.

3.4 System Design

The **Figure 3.2** illustrates the overall design blueprint of the Research Survey Assistant. The explanation is as follows:

- **On the top layer:** The top-down survey begins with a comprehensive bird-eyes view, offering condensed abstracts of over 100 papers in a ***fish-bone*** structure. This structure comprises nodes and edges that illustrate hierarchical and causal relationships. Each node contains a concise summary of fewer than 10 key phrases. The ***fish-bone***'s overall summary is derived from the issue ontology found in the papers' intro-

duction sections, created through issue clustering, classification, and prompt-engineering. Sentences containing issue ontology are extracted from the infrastructure data by extracting section information and their corresponding paragraphs.

- **In the middle layer:** Building on the direction provided by the bird-eyes view, the Longitudinal insight view offers condensed insight summaries from over 10 papers, presented as an Insight knowledge-tree. This tree comprises paper nodes connected by edges that summarize issue relevance between nodes. Each node and edge summary consists of sentences highlighting insights from Resolved and Finding issues in specific research papers. The Insight knowledge-tree's paper summaries stem from the issue ontology found in the papers' 'conclusion' and 'limitation' sections, created through Issue classification and prompt-engineering techniques. Sentences containing issue ontology are extracted from the infrastructure data using section information and their associated paragraphs.
- **At the bottom layer:** As an extension on the Longitudinal insight view, the Cross-sectional insight view offers a diff-table visualization for novice researchers seeking in-depth exploration of internal articles. This approach generates concrete summaries that reflect expert consensus viewpoints on the research topic. The Diff-table's paper summaries are derived from the full text of each paper. The process involves keyword-based extractive summarization to pinpoint relevant paragraphs, followed by prompt-engineering to create abstractive summaries.
- **Resource:** The views are generated using data derived from the secondary development of infrastructure data built by **CTBR** and *S2orc*. **Figure 3.4** provides a detailed description of this data-flow.

3.4.1 Front-end

To provide novice researchers with easily understandable multiple views for top-down survey logic and enable the system's back-end to generate corresponding view content based on user interactions, we designed a set of interaction logic between the front-end, user, and back-end. This design is illustrated in the **Figure 3.3**. The user experience begins with viewing a initial ***fish-bone*** diagram of a specific research topic. This diagram illustrates the research tasks within the topic and summarizes the research issues for each task. The process then unfolds as follows:

1. **User Interaction (US1)**: The user read the ***fish-bone*** diagram and selects a task and its corresponding research issues for in-depth

exploration.

2. **Back-end Processing (SS1)**: The system detects the user's **US1** action and initiates **SS1** operation. It retrieves relevant research papers and abstracts for longitudinal insight processing.
3. **Front-end Display**: The system generates and displays a longitudinal insight tree to the user.
4. **User Interaction (US2)**: The user studies this knowledge tree, identifies multiple papers for further investigation, and clicks on these papers to initiate **US2** operation.
5. **Back-end Processing (SS2)**: Upon detecting the user's **US2** action, the system executes **SS2** operation. It conducts cross-sectional insight processing on the user-specified papers.
6. **Front-end Display**: The system generates and presents a diff-table to the user.
7. **In-depth Analysis**: The user can now engage in comprehensive reading and analysis of multiple papers, facilitated by the structured information provided in the diff-table.

This multi-step process enables users to progressively deepen their understanding of the research topic, moving from a broad overview to detailed cross-paper comparisons.

3.4.2 Back-end

To implement the comprehensive top-down survey assistance process, the back-end requires a series of complex data flow processing steps based on infrastructure data. These include task clustering, issue classification, keyword scanning, and prompt engineering to generate corresponding summaries. Performing these processes synchronously with the front-end UI operations would cause significant delays due to the massive computational load. To address this, we preload and store processed data from the Fish-bone view and summaries generated for each paper in the research topic in the back-end before executing the interface. During system operation, it only needs to filter the back-end data according to user operations using appropriate algorithms, then display it in the front-end. This approach ensures a smoother user experience with minimal UI response time. The detailed back-end data logic is illustrated in the **Figure 3.4**. **Chapter 5 - 7** will provide detailed descriptions of the work-flows for **SS0**, **SS1**, and **SS2**.

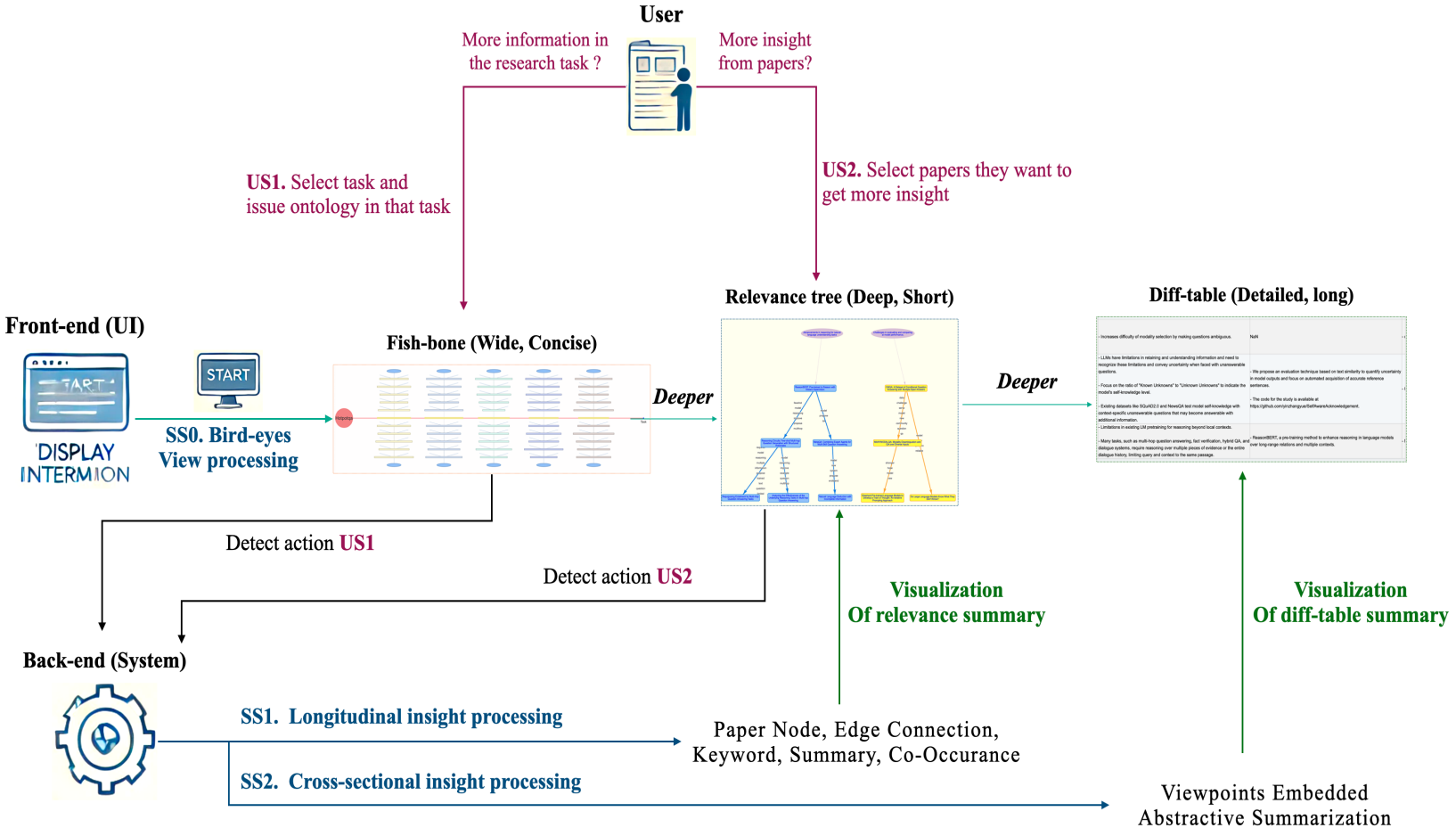


Figure 3.3: Front-end process

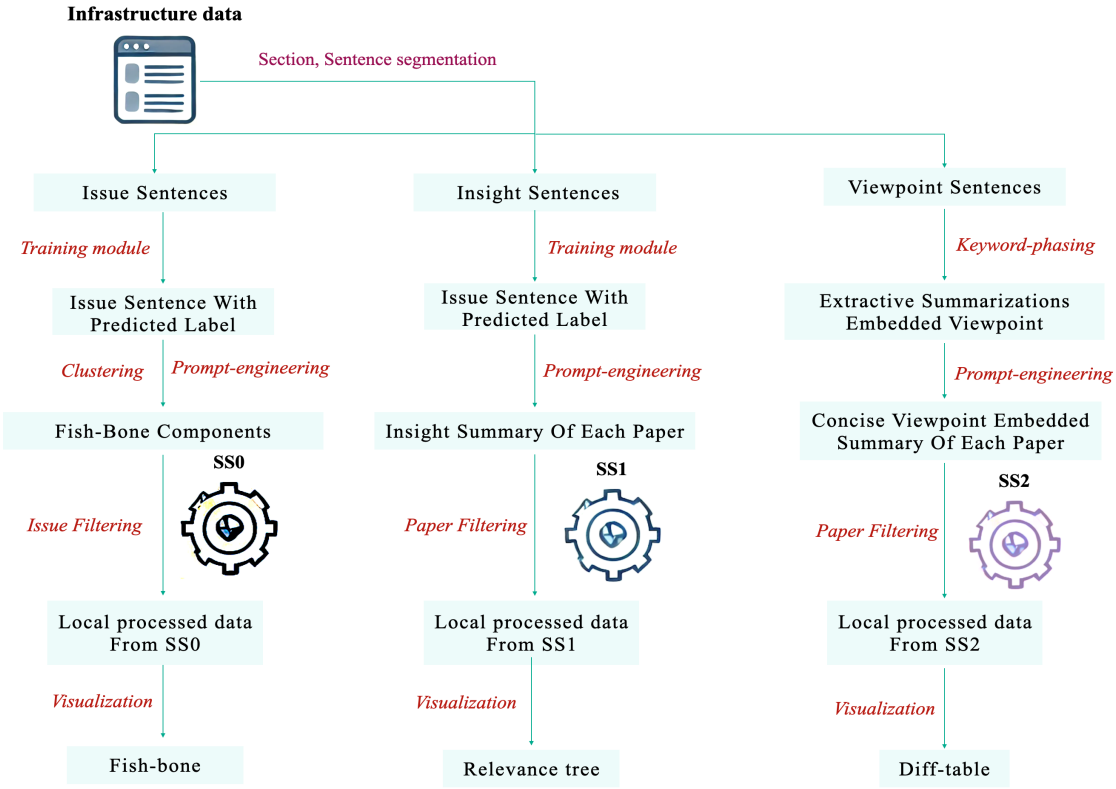


Figure 3.4: Back-end: Data-flow

Chapter 4

Infrastructure data building

4.1 Motivation

This chapter introduces the method of building infrastructure data for the survey assist interface. We need to extract and identify the body text within research paper *PDFs*, along with the paragraphs and their corresponding section titles, to construct layered infrastructure data that only carries linguistic information. However, identifying this linguistic information in actual paper *PDFs* is still challenging. Existing datasets, such as *S2orc* [21] and *Unarxiv* [77], sometimes contain information from figures, tables, and formulas. This leads to noise in the continuous paragraphs and causes inconsistencies in the sentences within the data. The issue arises because non-linguistic information often appears between paragraphs of linguistic information in articles. If these special objects in the papers aren't clearly distinguished, the generated infrastructure data will be impure. This impurity negatively impacts subsequent processes such as sentence analysis carrying issue ontology, summary generation, and summary visualization. The remaining problems of previous work and our corresponding solutions are shown in **Table 4.1**. Thus, we are dedicated to create a framework for document layout analysis using text block and compartment analysis to refine document layout to generate a cleaner infrastructure data.

4.2 Methodology-Definition

4.2.1 Overview

We are dedicated to create a framework for document layout analysis using text block and compartment analysis [79]. Our approach utilizes rule-based algorithms to process text within identifiable text blocks to implement rough compartment segmentation. Additionally, we employ machine learning models for multi-modal text block classification. Furthermore, by combining the classification result with rough compartment segmentation, we use a so-

Table 4.1: Remaining Problems & Our Solution

Remaining Problems	**Our Solution**
When text with features similar to body-text appears inside a figure or table, it becomes difficult to distinguish the figure or table region.	We analyze the characteristics of the text in figures/tables, vectorize and classify them in order to distinguish them from the body-text.
Large-scale pre-training data may not yield good object recognition results in specific scientific document formatting.	To achieve better object recognition results, we utilize small-scale data specifically formatted for scientific documents.
Sometimes, machine learning methods may incorrectly identify simple elements in a document	To further reduce misclassification in machine learning, we incorporate rule-based methods to create sophisticated compartment segmentation.

phisticated compartment refinement algorithm to achieve object recognition. Our framework consists of three stages, as illustrated in **Figure 4.1**.

- The first stage is preprocessing, which provides detailed instructions on parsing *PDFs* and extracting information from text blocks from scientific documents.
- The second stage involves text block classification, using a combination of rule-based methods to identify single-modal text blocks and machine learning to classify complex text blocks.
- The third stage focuses on the algorithm of compartment segmentation and object recognition for figures and tables based on the results of text block classification in the second stage.

4.2.2 Internal Environment of scientific document

To gain a deeper understanding of document layout, it is important to define the Internal Environment in the document for developing a simulation model, similar to urban planning or apartment room layout design. In this study, we categorize the internal structure of scientific documents into three levels. In this section, we provide specific definitions for the base domains, compartments, and text blocks within the context of scientific documents, as well as the roles they play in our overall system. The intuitive hierarchical

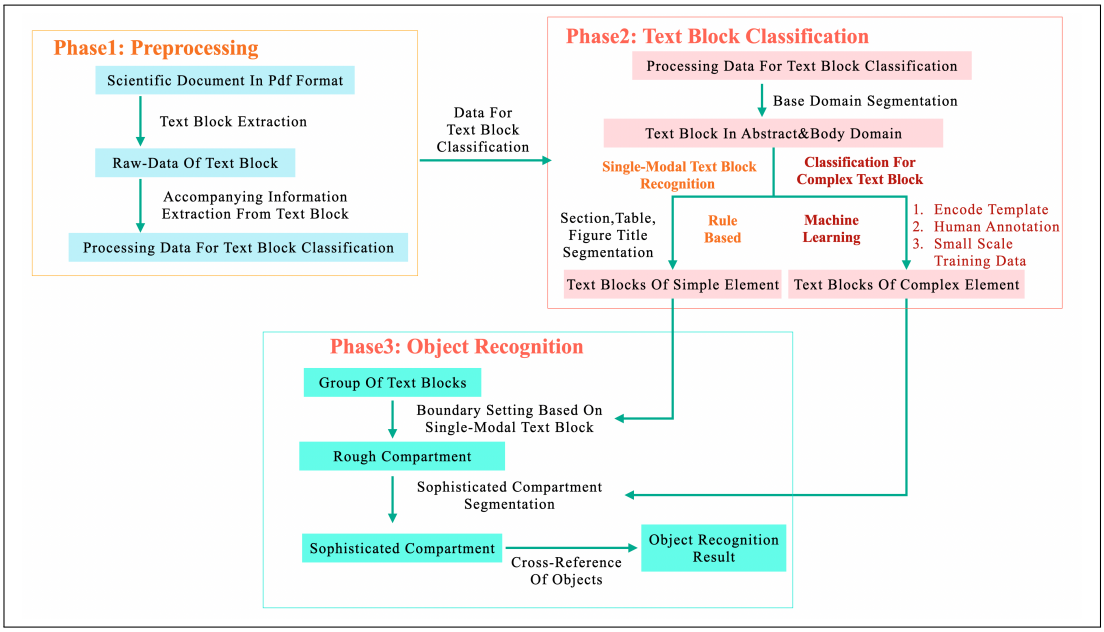


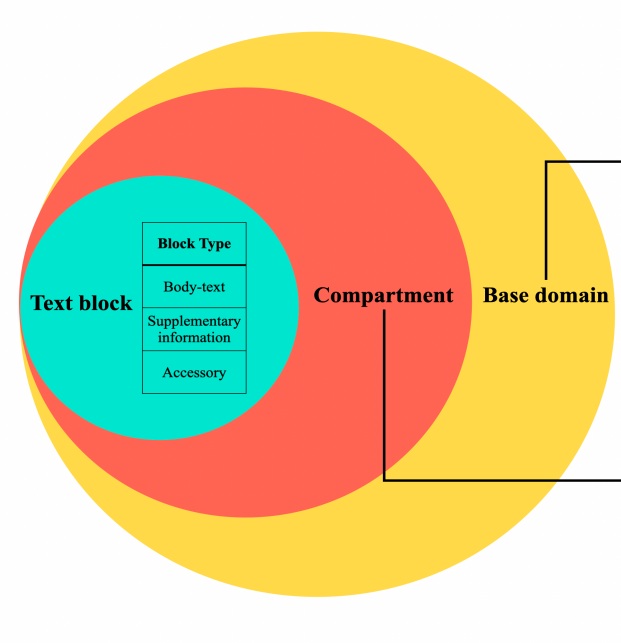
Figure 4.1: Overview

structure is shown in **Figure 4.2**.

- 1. The first level consists of base domains that form the overall structure, such as basic information domain (the region before the 'chapter 1 in paper' included title, author information, mail address, abstract.etc.), Its function is to provide readers about the meta information of this article and a general understanding. Comparable to the entrance area inside a house, when you step into the entrance area of a house, you may see the overall style and structure of the house, such as the color scheme, the general layout of the rooms, and the decorative style. This will give you a rough impression of the house.

- 2. The second level further subdivides each base domain into compartments, representing the different functions they serve. For instance, a model diagram can provide readers with an overview of research methods, helping them organize their thoughts and injecting energy into their understanding of the documents. Comparable to the dining compartment , which provide people with the necessary energy for their daily lives that promote engagement in more activities within the house.

- 3. The third level comprises text block within the compartments, which



Object	Base-Domain	Room area (Analogy)
Title, Author Information, Abstract	Basic information	Entrance area
Body-text, figure, title…	Body	Living area
Reference list	Reference	Rest area
Contents in Appendix	Appendix	Working area

Object	Occupy	Room Compartment (Analogy)
Section	Boundary	Door
Figure& Table title	Boundary	Door
Figure& Table Content	Compartment	Gaming Compartment
Footnote	Compartment	Dining Compartment
Equation	Compartment	Fitness Compartment

Figure 4.2: Internal Environment of scientific document

are the components that make up the compartments. Each text block also has specific functions that contribute to the overall function of the compartment. For instance, in a table, some text blocks contain data of accuracy rate that gives readers an intuitive expression to help them understand the intended purpose of the table as conveyed by the author. Comparable to the furniture in the dining compartment, such as dining tables, chairs, and wine cabinets, these furniture items are physical objects that we use in our dining process. They provide us a more tangible experience of the dining scene.

4.2.2.1 Text block in scientific document

The text block is a collection of different character strings. Just like authors organize scientific documents, they often group strings that express specific contents for formatting purposes. When these strings are closely arranged, they tend to form groups. Some text blocks feature single-modal expressions that convey singular and easily processed information, such as figure titles, table titles and section titles. These can be classified using rule-based and regular expression methods. However, rule-based methods do not easily understand some text blocks, particularly those embedded in figures/tables

and separated from the body-text. These multi-modal text blocks appearing in tables may represent measurements of experimental evaluation criteria, while those appearing in figures may describe components of research modules. These text blocks serve different purposes than similar-looking text blocks in the body-text. In this study, we aim to use machine learning methods for feature recognition of multi-modal text blocks. The types of text blocks are summarized in **Table 4.2**. To enable machines to analyze text blocks containing multi-modal information, we categorized the usage of text blocks in scientific documents that carry multi-modal information into three categories: Body-text, Supplementary information, and Accessory information [78]. Their definitions are shown in **Table 4.3**.

Table 4.2: Type & characteristics of Object

Object	Type	Reason	Method
Section Title	Obvious	• Continuity of section numbers • Specific fonts for section titles	RB
Tab&Fig Title	Obvious	The format of Tab/Fig titles is generally consistent within the same academic publication.	RB
Body-Text	Unobvious	Discontinuous data	ML
Figure	Unobvious	Irregular text block included	ML
Table	Unobvious	Irregular text block included	ML
Page_Num	Unobvious	Similar text block in tables	ML
Footnote	Unobvious	Similar text block in body-text	ML

Table 4.3: Type of text block

Type	Disciption
Body-text	Sentence group in body of article
Supplementary information	Figure and Table regions, Figure and Table titles Equations, Algorithms, Sections title
Accessory information	Page number, Footnote, Running title, Meta information

4.2.3 Definition of Compartment

This study defines a compartment in scientific documents as a group of multiple text blocks. A sample of the compartment is shown as **Figure**

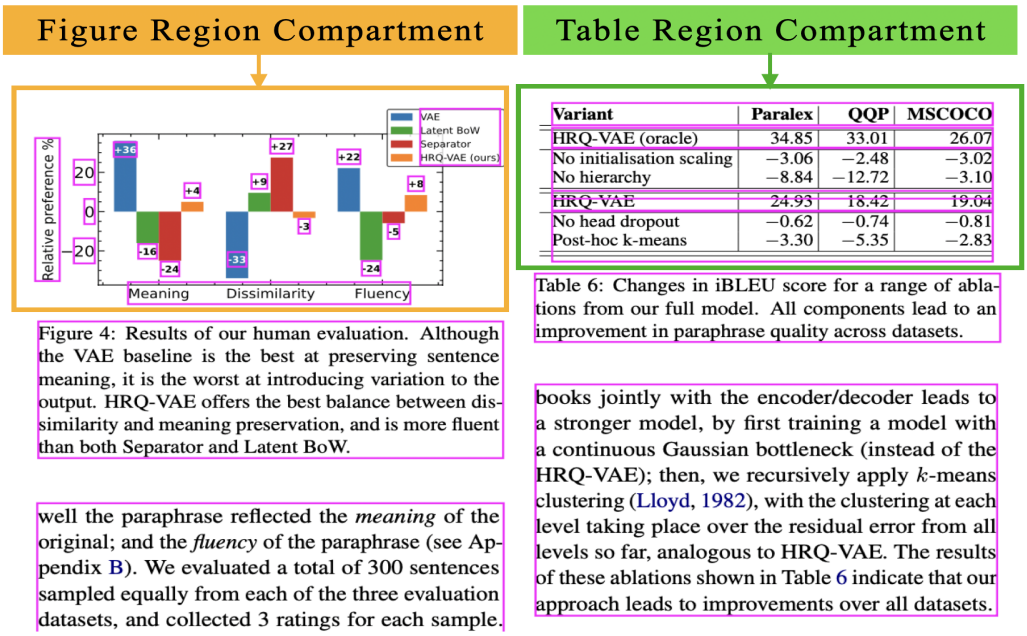


Figure 4.3: Figure & Table Compartment: Sample article [81].

4.3. These compartments provide a richer information combination than a single text block. For instance, when we see a figure in an article, we need to integrate the relationships between different parts of the figure to understand them fully. In this case, the figure region can be viewed as a compartment, where the information of different parts is often presented through text blocks or graphic elements. Therefore, accurately segmenting and recognizing the content inside compartments and the information they convey is crucial for improving machine perception of the document layout. In **Section 4.4**, we explain in detail how to utilize the results of text block classification for object detection based on a compartment segmentation algorithm.

4.3 Methodology-implementation

4.3.1 Phase1:Preprocessing

4.3.1.1 Text block extraction

After 2000s, the *PDF* format gradually gained popularity in academic publishing. Simultaneously, pre-print servers storing *PDF*-formatted papers emerged, accelerating the dissemination of research results [80]. In recent

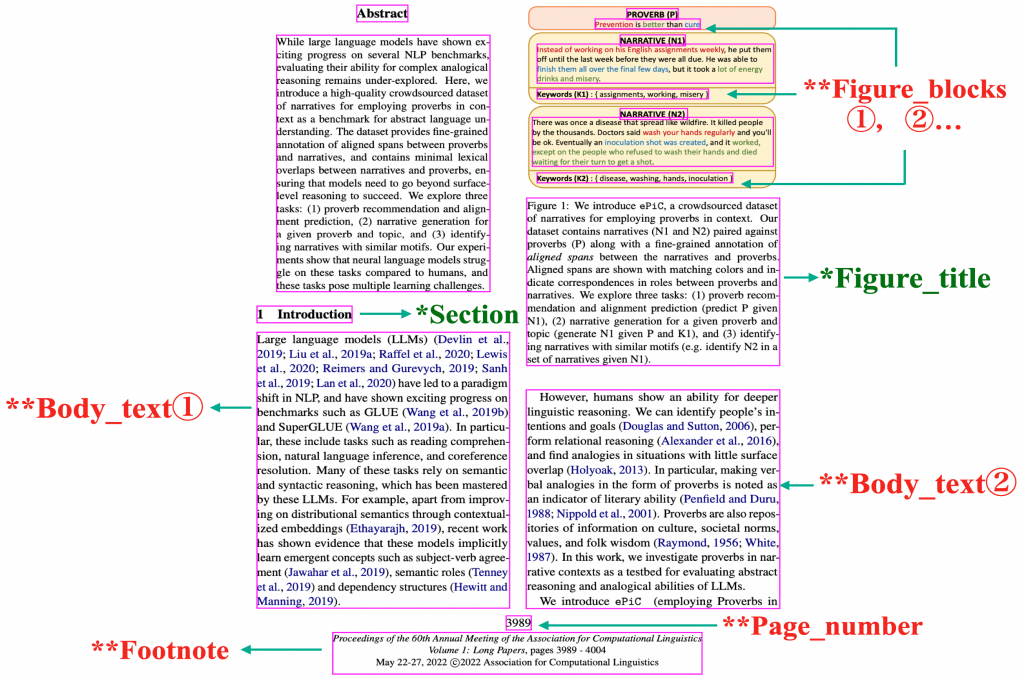


Figure 4.4: Unstructured page layout: Sample article [97].

years, *PDF* has become the near-universal format for scientific document. As a fundamental element of our **CTBR** framework - Text blocks, we first need to perform text block extraction based on obtaining the text grouping, line spacing, and column spacing in scientific *pdf* documents. This process can be automated by using the external library *pymupdf* [82] [78] in Python. By analyzing the underlying structure of scientific documents, such as line and column spacing, certain characteristics can be used to divide the text of a *PDF* file into multiple text groups. These text groups are distributed independently. We extract their bounding boxes to form our fundamental element - text blocks that contain specific content such as body-text, table headers, axis scales .etc. The combination results of text blocks compose the unstructured page layout, as depicted in **Figure** 4.4. The *Page.get_text("blocks")* method of *pymupdf* can be adopted to extract the text blocks from each page [82].

4.3.1.2 Accompanying information extraction from text block

When using *pymupdf* to extract text blocks, we can also obtain the accompanying text's information, such as font type, size, style [82]. Providing this

information can assist the machine clearly determine the types of information within the text blocks, makes it possible to optimize the extraction process and improve the accuracy of extracting specific features to execute text block classification . For example, the font size can be used to identify main sections of the document or differentiate between headings and body text. This assists us in setting boundaries for our subsequent compartment segmentation work, resulting in easily obtainable and low-noise output [78]. The accompanying information extracted in this study is shown in **Table 4.4**.

4.3.2 Phase 2-1: Rule-based Implementation for simple text block element

4.3.2.1 Base domain segmentation

We defined the Basic Information Domain, Body Domain, Reference Domain, and Appendix Domain as the fundamental domains. The subsequent work of this research specifically focuses on the Body Domain [78]. We utilize regular expressions and text accompanying information within the text blocks to perform compartment segmentation. The segmentation method is illustrated in **Table 4.5**.

4.3.2.2 Single-modal text block recognition

This study focuses on processing scientific documents that have specific section numbers assigned. There are two types of section titles: main-section and sub-section. **Table 4.6** shows each matching method using regular expression. The font characteristics(font type and font size) of the text are also utilized to distinguish similar patterns in the body-text. Figure and table titles use a similar recognition method, as shown in **Table 4.6**.

4.3.3 Phase 2-2: Classification for complex text block element

In contrast to the processing method described in the previous section for single-modal text blocks, effectively classifying the information contained in multi-modal text blocks is challenging using rule-based algorithms. This is because the conditions required for rule-based processing are complex, making it difficult to capture individual cases. In this study, the difference and definitions of multi-modal text block and single-modal text block are shown in **Figure 4.5**. This chapter extracts features from multi-modal

Table 4.4: Accompanying information and their usage

Accompanying information	**Usage**
Bounding box	Encode for left, right, top, bottom, width,height
Font size	• Boundary setting • Encode for font size, text blockdensity
Font type	• Boundary setting • Encode for font type
Media box	Compartment segmentation

Table 4.5: Base Domain segmentation

Domain	**Segmentation method**
Basic information	Before the section **'Introduction'**
Body Domain	• Regular expression to match **'Introduction':** `^Introduction\s*$` • Domain between section **'Introduction'** and **'Reference'**
Reference	• Regular expression to match **'Reference':** `^Referecnces\s*$`, • Domain between section**'Reference'** and **'Appendix'**

Table 4.6: Regular expression for single-modal text block recognition in 60th *ACL* Annual Meeting format

Element	Regular expression
Main section title	`^[0-9]{1,2}\s*[\.|,].*$`
Sub section title	`^[0-9]{1,2}(\.[0-9]{1,2}){1,4}\s+.*$`
Figure title	`^[F|f][I|i][G|g][U|u][R|r][E|e]\s*\d+\s*:.*$`
Table title	`^[T|t][A|a][B|b][L|l][E|e]\d+\s*:.*$`

text blocks and encodes them to address the issue of classifying multi-modal information. The **SVM** classifier in machine learning is then used to recognize these feature patterns for comprehensive classification.

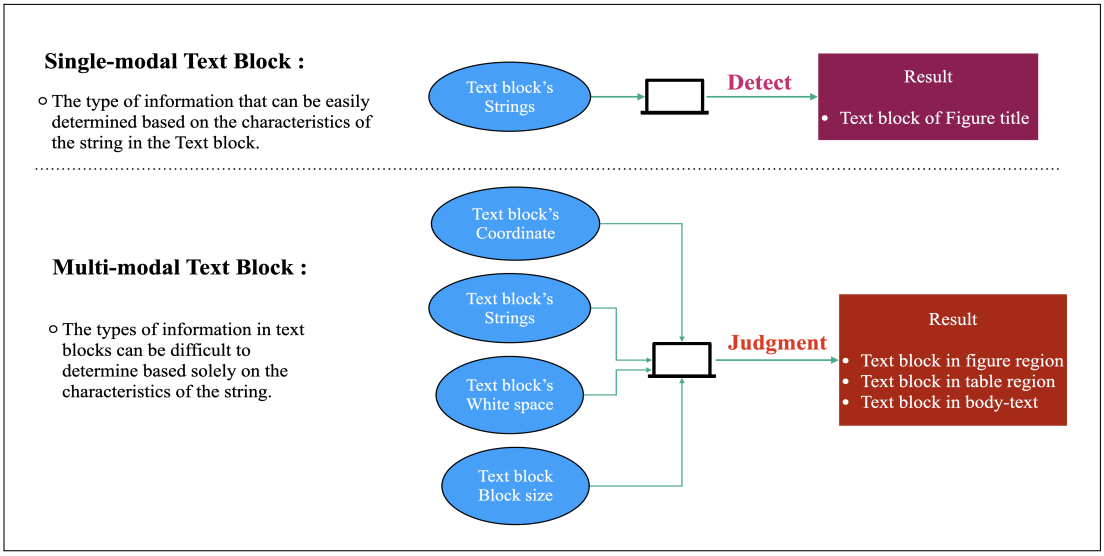


Figure 4.5: multi-modal text block and single-modal text block

4.3.3.1 Encoder template

To classify multi-modal text blocks into three categories (body-text, supplementary, accessory) using machine learning, we encode features for the text blocks into vectors and input them into the machine learning model. An encoder template is constructed based on accompanying text's information explained in **Section** 4.3.1.2, which directly processes the input of *PDF* documents to obtain the vectors representation in specific layout of scientific document. According to the characteristics of the arrangement of various objects on the pages of academic publications, as well as the intrinsic

structural features of objects, we decide nine vector components are combined to form a vector that represents the characteristics of text blocks. Each component of the vector is designed to incorporate the characteristics of corresponding text block. Specifically, the layout position of the text block on the page is determined by four components: left, right, top, and bottom positions. The size characteristics of the text block are represented by two components: width and height. The font type and font size are two components that characterize the text within the text block. Lastly, the characteristic of text density, when combined with the size of the text block and internal text features, represents the density of the text block structure. The encoding approach for each component is as follows [78]:

1-2, Left_Position and Right_Position: In the case of a body-text block that is often justified alignment, the goal is to accurately determine the starting position of each block and identify common characteristics of the bounding box. Specifically, we define the left/right-aligned as a boundary line , calculate the difference between the left/right coordinates of each block and the boundary coordinate.

$$Code_{(left|right)} = \frac{Block\_coordinate_{(Left|Right)}}{Boundary\_coordinate_{(Left|Right)}} \tag{4.1}$$

3-4, Top_Position and Bottom_Position: Footnote, page number and running title information is typically are typically located at the top(header), bottom, or corners of a page. Therefore, it is necessary to encode the text block of the footnote, page number, and running title to differentiate it from other content based on their positions on the page. The encoding method is the same as the one described in left/right Position but switches from left/right to top/bottom.

$$Code_{(top|bottom)} = \frac{Block\_coordinate_{(top|bottom)}}{Boundary\_coordinate_{(top|bottom)}} \tag{4.2}$$

By utilizing the encoding of top/bottom position, in conjunction with the left/right coordinates of the text block and font characteristics, we can enhance the recognizably of footnotes, page numbers, and running title information.

5-6, Width_Length and Height_Length: While the blocks of body-text are distributed more regularly, it is necessary to catch the width and height characteristics of the blocks to recognize the supplement information blocks, as they are always irregularly distributed in figure/table region. Therefore, in a document, set the largest width or height as the standard. Then, perform a division of the width/height of each block accordingly and

encode it.

$$Code_{(width|height)} = \frac{Block\_size_{(width|height)}}{Max\_size_{(width|height)}} \tag{4.3}$$

7, Font_Type(ft): In most cases, the font used for the body-text has the highest occurrence in scientific documents. Therefore, according to the font acquisition method in **Section 4.3.1.2**, all text blocks could be scanned to calculate the total number of characters for each font. The font type that appears most frequently is then identified as the body-text font. Next, the most frequently appearing font in each block is calculated, and if it is the body-text font, encoded as **'1'**. Otherwise, it is encoded as **'0'**.

$$Body_{font} = Index\_of\{Max\{\sum ft1, ..., \sum ftN\}\} \tag{4.4}$$

$$Code_{ft} = \begin{cases} \mathbf{1}\ (Body\_font) \\ \mathbf{0}\ (Others) \end{cases} \tag{4.5}$$

8, Font_Size(fs): First, we can determine the corresponding font size for the body-text based on its font type, and then set it to the standard font size, which is the most commonly used font size in a scientific document. Next, identify the most prevalent font size in each text block and assign it as the font size for that block. Lastly, calculate the ratio of each text block's font size to the standard font size and encode it. In equation (4.6 - 4.8), **fss** means font size in that text block.

$$Body_{fs} = Index\_of\{Max\{\sum f_s1, ..., \sum f_sN\}\} \tag{4.6}$$

$$Block_{fs} = Index\_of\{Max\{\sum f_{ss}1, ..., \sum f_{ss}N\}\} \tag{4.7}$$

$$Code_{fs} = \frac{Block_{fs}}{Body_{fs}} \tag{4.8}$$

9, Density of text block: In contrast to the basic characteristics of a text block mentioned earlier, we propose the concept of a higher-dimensional feature of text block density. This feature is defined as the ratio of the occupied area of text within a text block to the size of the text block. A larger ratio indicates a smaller blank area within the text block. Typically, body-text containing intensive text arrangement has a small occupied area of blank

space. On the other hand, text blocks corresponding to table information tend to have a larger blank area due to the presence of spaces and empty lines. Therefore, we calculate the text block density using equation (4.9 - 4.11), which combines several characteristics mentioned in the previous section to determine the category of a text block.

$$Code_{density} = \frac{Area\ of\ text\_block}{Area\ of\ text\ occupy} = \frac{S_{block}}{Length_{text} * Code_{fs}} \tag{4.9}$$

$$S_{block} = Width\ of\ Block_{size} * Height\ of\ Block_{size} \tag{4.10}$$

$$Length_{text} = Count\ of\ text\ in\ block \tag{4.11}$$

4.3.3.2 Annotation & Classifier for text block

To enable the machine to recognize the types of information mentioned in **Section 4.2.2.1** for each text block, we follow a process of encoding and vectorizing the text blocks. These text blocks, which contain multi-modal information, are manually annotated by humans. By combining the vectors of the constructed text blocks, we create a dataset for text block classification [78]. This dataset has short annotation time, easy judgment, and high accuracy. An example of the labels can be seen in **Figure 4.6**. A researcher from computer science can annotated 10 *PDF* documents for multi-modal text block element, following the object description in **Table 4.2 - 4.3** and **Figure 4.4**, in just 45 minutes. The single-modal text block element are automatically annotated using the rule-based method detailed in **Section 4.3.2**. The training dataset for multi-modal text block classification is constructed by combining the vectors of the text blocks with the corresponding human annotations. Next, we use a Support Vector Machine **(SVM)** classifier to classify the type of text blocks. We chose the **SVM** classifier for the following reasons: **SVM** is a highly accurate classifier in machine learning that maximizes the margin to improve classification accuracy for unknown data [84]. In addition, The **SVM** classifier's margin and tolerance range setting are adaptable and versatile for scientific documents with complex structures, such as special cases of text blocks in figures with the same font and size as the body-text [78]. Setting margins and tolerances enables us to distinguish prominent characteristics of text blocks while minimizing overfitting as much as possible. Hence, we firmly believe that **SVM** can achieve high-precision classification on our small-scale dataset due to these features.

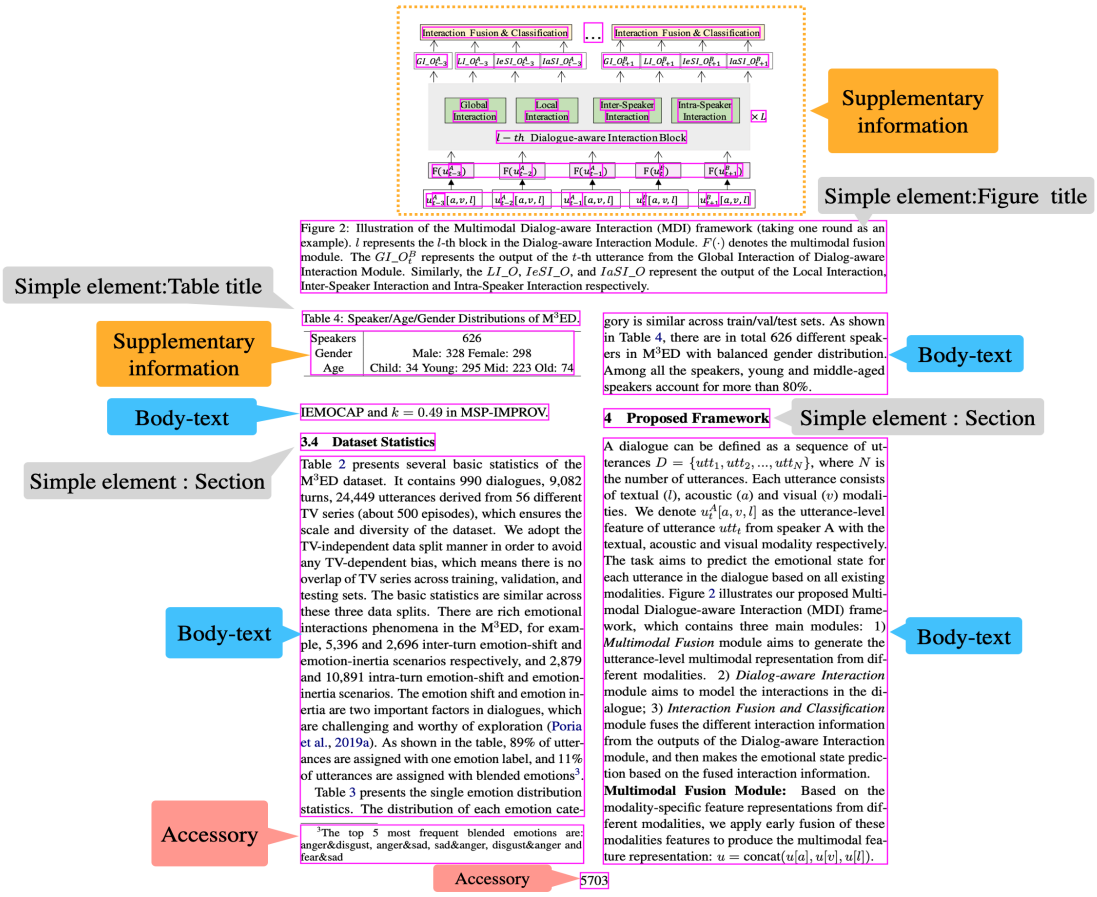


Figure 4.6: Human annotation for text blocks [83].

4.3.4 Phase3 : Compartment Segmentation & Object Recognition

Using the text block classification results from Phase 2, we aim to cluster the classification results of text blocks into group levels for compartment segmentation and object recognition. Initially, we use single-modal text blocks such as section titles, figure titles, and table titles to establish the boundaries (refer to **Figure 4.7**). The content between these single-modal text blocks (for double-column layout documents, page break coordinates need to be set) may represent a compartment containing body-text or a compartment containing figures and tables. By combining the classification results of multi-modal text blocks, we statistically determine the category of the text blocks within this compartment and identify the type of object they represented. Finally, we establish the correlation between figure/table titles and their respective compartment based on their positions.

4.3.4.1 Boundary setting based on simple text block

We analogize the boundary lines in the document to doors or walls separating independent rooms. These boundary lines divide the entire document into several rough compartments. We determine the boundaries by considering the position and order of simple text blocks in the document, as illustrated in **Figure 4.7**.

- **Type1: Figure & Table Title Crossing the Central Axis**

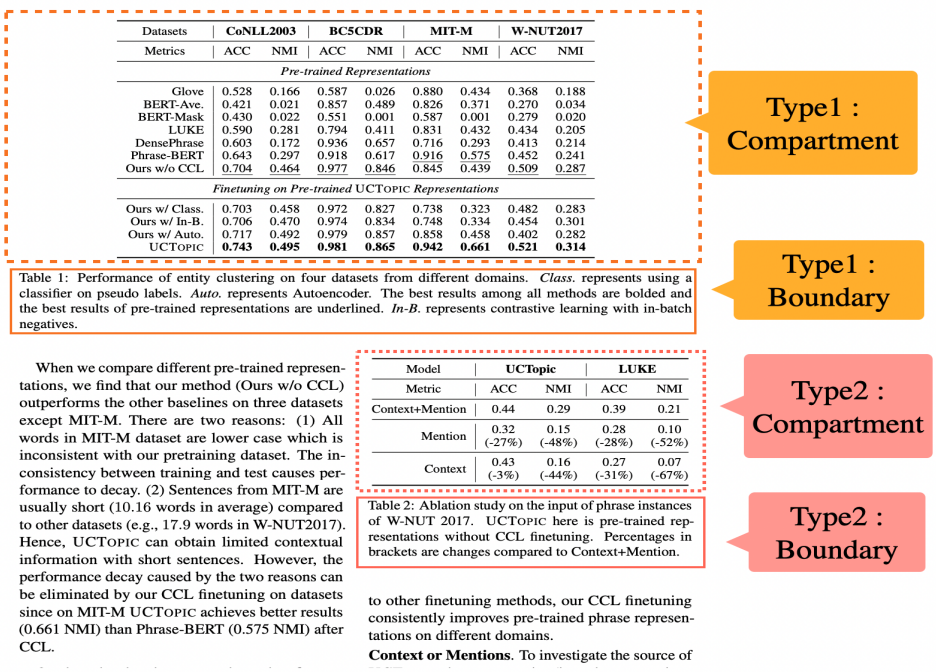
If the text block area corresponding to the title of a figure or table crosses the central axis of the document, the region of the figure/table will also cross the central axis of the document.

- **Type2: Figure & Table Title Not Crossing the Central Axis**

Conversely, if the text block area corresponding to the title of a figure or table does not cross the central axis of the document, then the region of the figure or table will follow the dual-column layout.

4.3.4.2 Sophisticated Compartment Segmentation & Object recognition

In this section, we utilize the classification results of the text blocks and the position of the boundary to enhance the refinement of the rough compartment. We begin by assigning sequential numbers to the rough compartment and the boundary based on their arrangement in the document. Next, we iterate through the sequence of appearance of the figure and table titles



Datasets	CoNLL2003		BC5CDR		MIT-M		W-NUT2017	
Metrics	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
Pre-trained Representations								
Glove	0.528	0.166	0.587	0.026	0.880	0.434	0.368	0.188
BERT-Ave.	0.421	0.021	0.857	0.489	0.826	0.371	0.270	0.034
BERT-Mask	0.430	0.022	0.551	0.001	0.587	0.001	0.279	0.020
LUKE	0.590	0.281	0.794	0.411	0.831	0.432	0.434	0.205
DensePhrase	0.603	0.172	0.936	0.657	0.716	0.293	0.413	0.214
Phrase-BERT	0.643	0.297	0.918	0.617	0.916	0.575	0.452	0.241
Ours w/o CCL	0.704	0.464	0.977	0.846	0.845	0.439	0.509	0.287
Finetuning on Pre-trained UCTopic *Representations*								
Ours w/ Class.	0.703	0.458	0.972	0.827	0.738	0.323	0.482	0.283
Ours w/ In-B.	0.706	0.470	0.974	0.834	0.748	0.334	0.454	0.301
Ours w/ Auto.	0.717	0.492	0.979	0.857	0.858	0.458	0.402	0.282
UCTopic	**0.743**	**0.495**	**0.981**	**0.865**	**0.942**	**0.661**	**0.521**	**0.314**

Table 1: Performance of entity clustering on four datasets from different domains. *Class.* represents using a classifier on pseudo labels. *Auto.* represents Autoencoder. The best results among all methods are bolded and the best results of pre-trained representations are underlined. *In-B.* represents contrastive learning with in-batch negatives.

When we compare different pre-trained representations, we find that our method (Ours w/o CCL) outperforms the other baselines on three datasets except MIT-M. There are two reasons: (1) All words in MIT-M dataset are lower case which is inconsistent with our pretraining dataset. The inconsistency between training and test causes performance to decay. (2) Sentences from MIT-M are usually short (10.16 words in average) compared to other datasets (e.g., 17.9 words in W-NUT2017). Hence, UCTopic can obtain limited contextual information with short sentences. However, the performance decay caused by the two reasons can be eliminated by our CCL finetuning on datasets since on MIT-M UCTopic achieves better results (0.661 NMI) than Phrase-BERT (0.575 NMI) after CCL.

Model	UCTopic		LUKE	
Metric	ACC	NMI	ACC	NMI
Context+Mention	0.44	0.29	0.39	0.21
Mention	0.32 (-27%)	0.15 (-48%)	0.28 (-28%)	0.10 (-52%)
Context	0.43 (-3%)	0.16 (-44%)	0.27 (-31%)	0.07 (-67%)

Table 2: Ablation study on the input of phrase instances of W-NUT 2017. UCTopic here is pre-trained representations without CCL finetuning. Percentages in brackets are changes compared to Context+Mention.

to other finetuning methods, our CCL finetuning consistently improves pre-trained phrase representations on different domains.

Context or Mentions. To investigate the source of

Figure 4.7: Boundary setting & compartment [98].

that correspond to the boundary. Based on this strategy, we can infer the position and size of the compartments where the figures and tables are located, following the rules outlined below. This approach enables more precise compartment recognition.

1. If the boundary where the figure/table title is located is at the top of the page (without any compartments appearing before it), the compartment where the figure or table is located must be directly below the boundary.
2. If the boundary where the figure/table title is located is at the bottom of the page (without any compartments appearing after it), the compartment where the figure or table is located must be directly above the boundary.
3. If the boundary where the figure/table title is located is in the middle of the page (with compartments before and after it), we need to use the text block classification results from the rough compartment, following Equation (7.1 - 7.3), to infer whether the corresponding figure/table region appears directly above or below. In Equation (7.1 - 7.3), $\boldsymbol{S}$ means area.

$$Above|Below = Area\ of\ label\ 'Supplementary'\ Occupied \tag{7.1}$$

$$Above|Below = \frac{\sum S_{block\_label\ of\ 'Supplementary'}}{S_{Compartment}} \tag{7.2}$$

$$Compartment\ = \begin{cases} Above > Below\ ,\ Above\ is\ Figure/Table\ region \\ Above < Below\ ,\ Below\ is\ Figure/Table\ region \end{cases} \tag{7.3}$$

4. Once the boundary-compartment correspondence is confirmed, other boundaries cannot occupy the same compartment.
5. Since some compartments may contain additional information such as body-text or footnotes, we need to use the text block classification results for refining the rough compartment, following Equation (8.1 - 8.3) to further segment the figure & table region.

$$Bbox_{compartment} = (left_{pos},\ right_{pos},\ up_{pos},\ bottom_{pos}\ ) \tag{8.1}$$

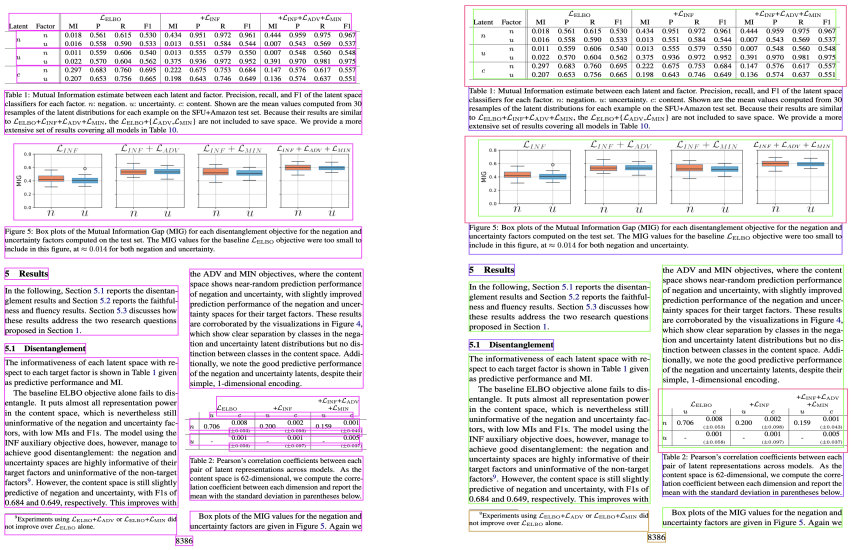
$$Left_{pos}|Top_{pos} = Min_{Left|Top} \sum Text\ block\ of\ 'supplementary' \tag{8.2}$$

$$Right_{pos}|Bottom_{pos} = Max_{Right|Bottom} \sum Text\ block\ of\ 'supplementary' \tag{8.3}$$

6. Referring to the pdffigure2.0 method [36], if the width of the figure/table compartment is larger than the figure/table title, it is determined as the figure/table compartment. If the block's width where the figure/table title is located is larger, resize the figure/table compartment to match that width. The results of the Compartment recognition samples are shown in **Figure 4.8**.

4.3.5 Applied processes and Usage areas of CTBR

In the previous sections, we detailed the development process of the **CTBR** framework. Its standout feature is the ability to achieve optimal object recognition results with a small dataset for specific scientific document typesetting. To adapt the **CTBR** for other typesetting, such as IEEE or ACM, follow the steps outlined below based on Figure 4.9:
(1) Use the method in **Section 4.3.1** to extract text blocks and their corresponding accompanying information for raw-data preprocessing.
(2) To create vectors for the dataset that a computer can recognize, vectorize each text block using the encoder-template provided in **Section 4.3.3.1**.



(a)Unaligned layout and text block (b)Figure,Table region detection

Figure 4.8: (b)Sample of Object Detection: Sample article [85]

(3) According to **Section 4.3.2**, use the rule-based method to identify and automatically annotate single-modal text block elements within the dataset. In the cases such as Roman numeral chapter numbers or non-IMRaD type articles [86], writing the corresponding regular expressions are requirement.
(4) Classify the multi-modal text block element using the human annotation method and classifier detailed in **Section 4.3.2**.
(5) Refer to the Compartment segmentation algorithm described in Section **4.3.4**, combining single-modal text block elements detected by rule-based results and multi-modal text block elements identified via machine learning results, to perform final object recognition.

4.4 Experiment

4.4.1 Scientific document collection in *PDF* format

We collected 768 articles in *PDF* format from the *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* to use as our experimental data. We designed text block classification and two-stage object recognition experiments to validate the effectiveness of the **CTBR** model in *60th ACL Annual Meeting* format.

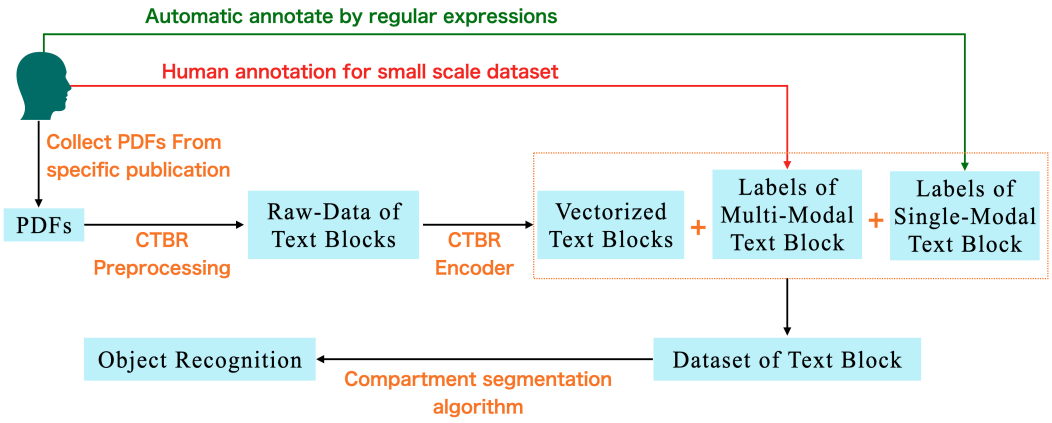


Figure 4.9: Applied Processes

4.4.2 Experiment(1): Text block classification

To test the learning performance of our small data, we randomly selected 10 articles from 768 research articles to build a training module. We encoded and annotated them with human input and constructed a dataset that took approximately an hour to complete [78]. For the data split, 90% were set as training data and 10% as validation data. The label distribution is shown in **Table 4.7**. Through a process of grid search [87] [88], we have determined the optimal training parameters for the **SVM** model, which are presented in **Table 4.8**. These parameters were selected based on a rigorous evaluation of various combinations of hyperparameters, including the kernel type, regularization parameter, and degree of the polynomial kernel. Our results demonstrate that these parameters significantly improve the performance of the **SVM** model, resulting in more accurate and reliable predictions.

Table 4.7: Details of training&validation dataset

Class	Number of labels	Ratio
Body-text	514	33.9%
Supplement	894	58.9%
Accessory	109	7.2%
Total	1518	-

Table 4.8: Details of SVM Parameter

Parameter	Value	Description	Degree
C	100	There are fewer misclassification points in the decision area.	Large
Gamma	0.1	The decision boundary is a simple decision boundary.	Small
Rbf	-	It can represent a non-linear boundary.	-

4.4.2.1 Evaluation : Text block classification

The training and validation dataset were randomly processed in 100 experiments to ensure the stability of the results. Each experiment was designed to minimize the influence of random variations and present a more precise depiction of the experiment outcome. Based on the experimental results conducted 100 times, we calculated the average value and statistics for the accuracy, precision, recall, and F1-score of each label, as presented in **Table 4.9**. We achieved an overall precision, recall and F1-score of over **95%** in classifying types of text blocks. Additionally, we achieved an AUC-score [89] close to 1, demonstrating a robust and effective classification model with an excellent balance between sensitivity and specificity. These results provide an acceptable basis for compartment segmentation in object recognition experiments.

Table 4.9: SVM Result of validation data

Result	All label	Body-text	Supplement	Accessory
Precision	0.974	0.935	0.987	0.938
Recall	0.954	0.980	0.952	0.991
F1-Score	0.963	0.957	0.964	0.963
AUC-Score	0.993	0.990	0.991	0.998

4.4.3 Experiment(2-1): Object Recognition for figure & table - Global Sampling

We used 20 randomly selected articles in *PDF* format as test data from remaining 758 documents(excluding the *PDF* documents used to build the training model) for object recognition. Initially, the sampling data in **Section 4.4.2** was used as input for the **SVM** classifier to build a training

module. After obtaining the classification result, a comparison experiment was conducted to verify the effectiveness of the Table and Figure recognition. The comparison was made with previous studies, including *VGT, Table Transformer, Table net, Pdffigure 2.0, and Tabula.*

4.4.3.1 Result

After finishing sophisticated compartment segmentation& processing, the object recognition accuracy was obtained. The comparison experiment results are shown in **Table 4.10**. We achieved competitive recognition results for figures and tables in specific academic publications with only a small amount of training data from 10 scientific documents.

4.4.4 Experiment(2-2): Object Recognition for figure & table - Manually Selection

In the previous section, we randomly selected 20 documents for the first phase of our global experiment. Some of these documents included pages with a simple layout - a single figure or table positioned at the top. To ensure that our model maintains a high recognition rate even with a complex page layout, we manually extracted 30 pages from 758 *PDF* documents with consecutive arrangements of figures/tables or the pattern of figures/tables within the body-text. These pages of *PDF* were used for subsequent experiments and analysis of error types. We selected high accuracy model in **Table 4.10** - *VGT, Table Transformer, and PDFfigure 2.0* as the comparison for our second stage experiment.

4.4.4.1 Result

The experimental results in **Table 4.11** show that for complex layout pages, the **CTBR** framework has a competitive edge over table transformer models in table recognition and a significant advantage over *VGT* in figure recognition. It also performs well compared to *pdffigure2.0*, although in some cases, the results are mixed with both wins and losses. We list the case study for analysis in **Section 4.4.4.2**.

CTBR excels particularly in accurately distinguishing text within figure regions from body-text. Moreover, with the refined rules implemented in sophisticated compartment segmentation, it can more accurately handle cases involving continuous figures/tables. As a result, it can correctly identify the figure/table titles and their corresponding regions in *60th ACL Annual Meeting* format.

Table 4.10: Object recognition result – Global Sampling 20 scientific documents(Number of Figure: 74, Number of Table: 95)

	Training Data Scale	**Accepted Figure**	**Accuracy Figure**	**Accepted Table**	**Accuracy Table**
Our approach	**10 documents of *60th ACL Annual Meeting format***	**70/74**	**0.945**	**90/95**	**0.947**
VGT (2023)	8868 document images	65/74	0.878	89/95	0.936
Table Tranformer (2022)	948K Tables in 100,000s articles	-	-	87/95	0.915
Table net (2020)	509 Images for scientific documents	-	-	56/95	0.589
Pdffigure2.0 (2016)	Rule-based	69/74	0.932	**90/95**	**0.947**
Tabula (2018)	Rule-based	-	-	22/95	0.236

Table 4.11: Object recognition result – Manually Selection 30pages of scientific document(Number of Figure: 39, Number of Table: 57)

	Accepted Figure	**Accuracy Figure**	**Accepted Table**	**Accuracy Table**
Our approach	**36/39**	**0.923**	**55/57**	**0.965**
VGT (2023)	28/39	0.718	**55/57**	**0.965**
Table Transformer (2022)	-	-	51/57	0.895
Pdffigure2.0 (2016)	**36/39**	**0.923**	54/57	0.947

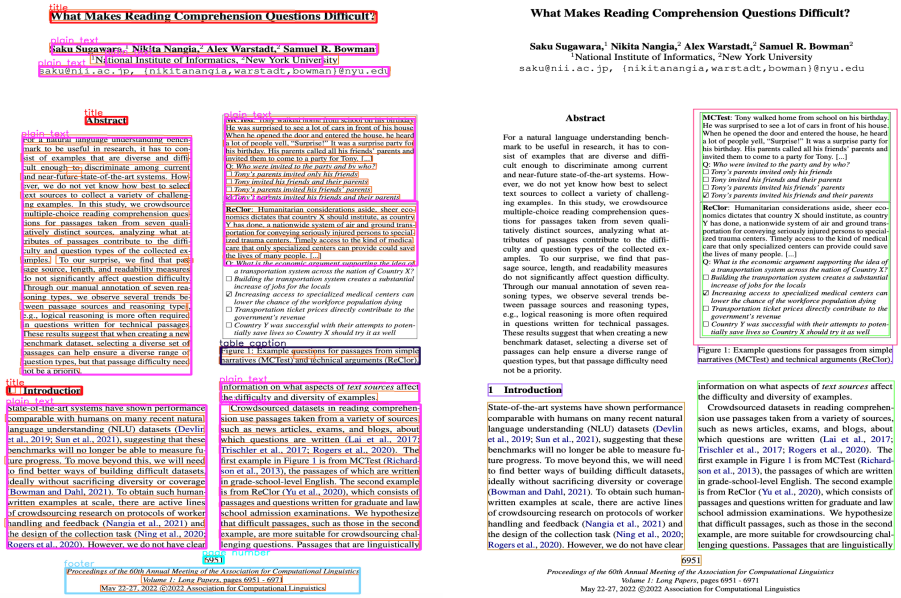
4.4.4.2 Comparison

Eventually, we conduct an analysis with high-performance models - *VGT, Table Transformer (TATR), and PDFfigure 2.0.*

- **Compare with VGT:** In the figure region, when there is information that is highly similar to the body-text, as shown in **Figure 4.10**, **CTBR** achieves more accurate recognition results in distinguishing the textual blocks of supplementary information within the figure. This is accomplished through the use of a sophisticated encoding template and accurate classification. Furthermore, as depicted in **Figure 4.11**, for side-by-side images, **CTBR** demonstrates excellent judgment regarding the relationship between the figure title and the corresponding figure region, based on sophisticated compartment segmentation.

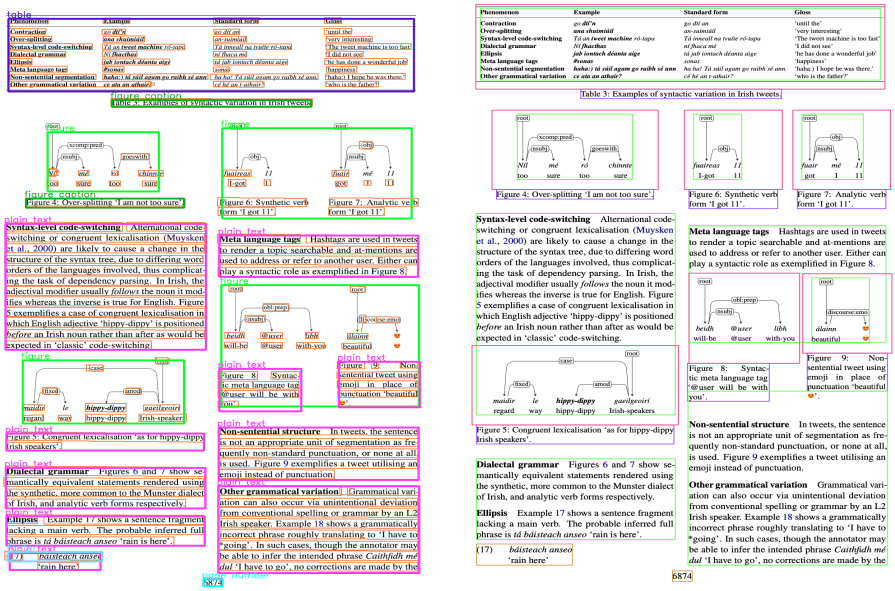
- **Compare with Table Transformer:** According to the experimental results, we have achieved higher recognition accuracy than the advanced module Table Transformer regarding table region recognition. As shown in **Figure 4.11**, our method demonstrated higher stability than complex pattern matching involving continuous tables on a page due to the clearly segmented boundary (table title).

- **Compare with Pdffigure 2.0:** The experiment showed that *pdffigure2.0* had trouble distinguishing between figure and body-text when multiple consecutive text blocks were in the image. We improved the accuracy of figure recognition by using font type and font size to differentiate text blocks from body-text, as shown in **Figure 4.13**. However, in some cases, *pdffigure2.0* has better edge recognition for



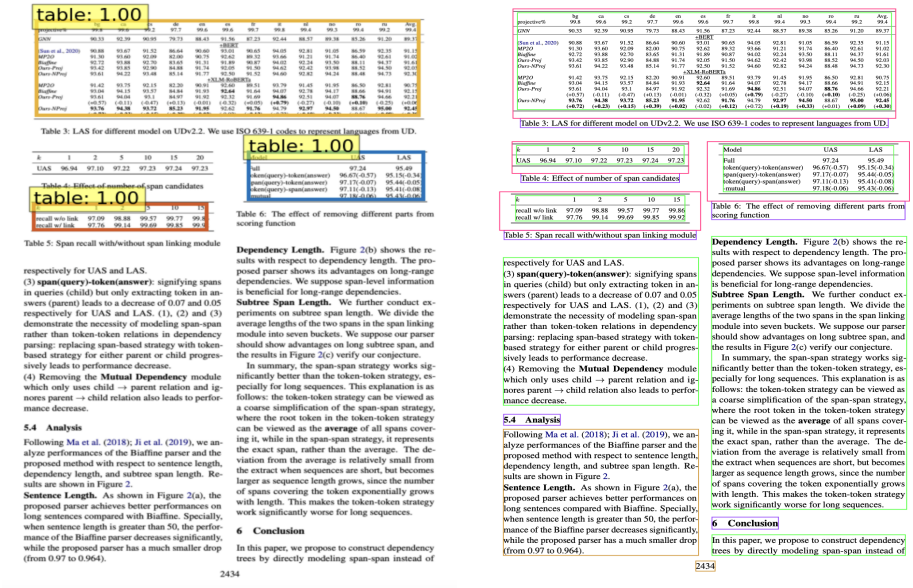
(a)VGT(Vision Grid Transformer) (b)Our approach

Figure 4.10: Sample of Comparing with $VGT(1)$: Sample article [90]



(a)VGT(Vision Grid Transformer) (b)Our approach

Figure 4.11: Sample of Comparing with $VGT(2)$: Sample article [91]



(a)Table_Transformer (b)Our approach

Figure 4.12: Sample of Comparing with *Table Transformer*: Sample article [92]

figure and table regions than **CTBR**. This is because *pdffigure2.0* has a sophisticated rule design for identifying whitespace areas on the page. In order to improve our **CTBR**, we need to not only encode the whitespace areas within the text block, but also establish rules for optimizing compartment segmentation based on the features of the whitespace areas on the page.

- **Improvement by CTBR in complex cases: Table 4.12** summarizes the advantages of **CTBR** in three complex cases.

Table 4.12: The advantages of **CTBR**

Pattern	**Improvement through CTBR**	**Reason**
Continuous multiple tables/figures	Reduce miss-detection	Establish clear boundaries and accurately identify the corresponding relationship between the figure/table title and the figure/table zone.
Figure filled with text	Prevent incomplete-coverage	Correctly distinguished between body-text and text inside figures and tables to achieve regional coverage optimization
Figure/Table sandwiched in the body-text	Reduce over-coverage	The encoding of density and compartment segmentation are used to clearly differentiate between body-text and figure / table zones. (due to the high text block density of body-text, while figures and tables are relatively sparse)

(a)Pdffigure2.0 (b)Our approach

Figure 4.13: Sample of Comparing with *Pdffigure2.0*: Sample article [93]

4.4.4.3 Error type of CTBR

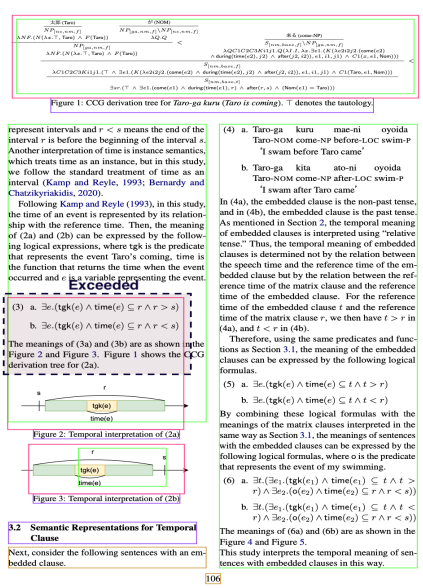
- **Out of scope:** Accurately identifying mathematical formulas that appear above the figure regions can be challenging when dealing with complex situations. This is because the symbols and elements in mathematical formulas sometimes have font types and font sizes that are very similar to those used in figures/tables. Additionally, there is a possibility that mathematical formulas appearing between body-text can disrupt the bounding box coordinates of the body-text's text block. As a result, the recognition may extend beyond the range of the figure region , as shown in **Figure 4.14(a)**. To address this issue, we realized the need for a more refined encoding and categorization of the text of formula within the body-text to differentiate it from other objects.

- **Incomplete coverage of figure regions:** In cases of a large number of text blocks within a figure, this study utilizes those text blocks for compartment segmentation and organization. However, there may be rare situations where the edges of the figure we recognized cannot fully cover the figure region, as shown in **Figure 4.14(b)**. Setting more precise boundaries at the page edges is recommended to address this issue. This approach can effectively improve the coverage of the figure

edges.

- **Incomplete coverage of table regions:** While this study successfully encoded the features of text block density to differentiate text blocks in tables that have the same font type and font size as the body-text, there are still some instances of incomplete coverage of table regions, as seen in **Figure 4.15(c)**. Hence, enhancing the compartment segmentation algorithm by establishing more specific rules for the typical positioning of figures and tables is necessary.

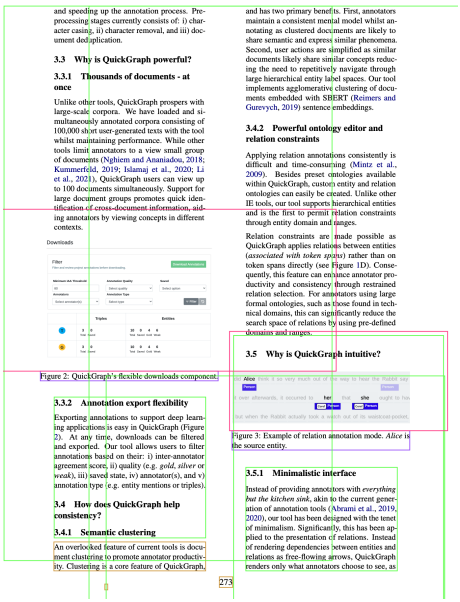
- **Text block parsing error:** During the text block phasing phase of this study, there is a minimal probability of encountering errors in extracting the coordinates of text blocks using the *pymupdf* API, as shown in **Figure 4.15(d)**. This can lead to deviations in the bounding box coordinates during the subsequent compartment segmentation, which may fail to recognize figures/tables.



(a)Out of scope:Sample article [94] (b)Incomplete coverage of figure regions [94]

Figure 4.14: Error type of **CTBR**-1

(c)Incomplete coverage of table regions [95]



(d)Text block parsing error [96]

Figure 4.15: Error type of **CTBR**-2

4.5 Summary

We developed an advanced framework called **CTBR** for text block classification and object recognition in scientific documents. This framework first defines the hierarchical structure of scientific document's layout, including base domains, compartments, and text blocks. We developed a bottom-up level encoded template to refine these text blocks, which contain multimodal data, and performed text block classification using machine learning and rule-based methods. Furthermore, we achieved more advanced compartment segmentation and object recognition using the classification results. The effectiveness of this framework was demonstrated through hierarchical scientific document layout analysis, using a small-scale training dataset and an **SVM** classifier for text block classification. We also developed a specialized algorithm for compartment segmentation to determine the region of figures and tables based on the classification results of text blocks, achieving an accuracy of over 90%. Overall, the experimental results showed the effectiveness of this framework.

Future tasks to improve this framework include the following:

- **1. More refined compartment planning algorithm:**

This study divides the sections of figures, body-text, and footnotes in scientific documents. One of the future challenges is to add a finer level of division within the body-text, such as recognizing compartments of equations, itemized forms, algorithm areas, and lists.

- **2. Compartment Internal Functional Differentiation**
  In this work, we acknowledge the importance of the compartment in scientific documents. However, the function within these compartments is also crucial for a comprehensive understanding of the document. For example, the text blocks in figures can be considered components, each with a special meaning. In a statistical graph, the data on the horizontal and vertical axes reflect the range of the object, while the names of the axes indicate the measured indicators and so on. Similarly, the components of a model diagram in a scientific document provide a clear representation of the input-output and basic logic of the model. Exploring such detailed information further can optimize the dataset for scientific document understanding and greatly enhance the interpretability of figures in scientific documents.

- **3. Dynamic programming and reinforcement learning-based methods for improving object recognition accuracy:**
  In this study, we classified the text blocks in the compartment into three categories and calculated the proportion of the text blocks occupied in each category to infer the characteristics of the compartment. Although we achieved satisfactory accuracy, there were some complex situations, as described in **Section** 4.4.4.2. These situations involved overlaps and incomplete recognition of compartments when determining the figure/table area to which the figure/table title belongs. To improve recognition accuracy, we can design a simulation model based on the document environment described in **Section** 4.2.2. This design will focus on dynamic programming and reinforcement learning. For example, we can create a reward function based on the features of the text blocks. This function will convert the probabilities obtained from machine learning classification results into rewards corresponding to each label. An agent can also be adopted to explore the document environment and learn the optimal compartmentalization pattern.

4.6 Infrastructure data Building for survey assistance interface

In text block classification, we identified and extracted the approximate areas of body text, then applied the **CTBR** framework to precisely distinguish non-linguistic information—such as figures and tables—interspersed within the body-text. This series of operations helped us extract more continuous and pure body-text. In addition to *60th ACL Annual Meeting* documents, we annotated a small sample of text blocks from prominent publications like *IEEE* and *ACM*. We then refined the body text using the **CTBR** framework to enhance the diversity of infrastructure data. Subsequently, we combined the section titles with the corresponding positions of paragraphs in the body text to establish hierarchical relationships among section titles, and individual paragraphs in body-text, thereby composing the infrastructure data. In the research survey assistant interface, we need to split the infrastructure data by section. The body text belonging to specific sections forms the dataset in the subsystem, with detailed divisions as follows:

- **'Introduction' part for bird-eyes view:** This part of dataset comprises body-text from the Introduction sections of papers. These sections typically outline the general direction of the research topic, guiding readers by providing a broad knowledge background and conveying the author's proposed logic. By combining Introduction part from multiple articles, novice researchers may gain a comprehensive overview of the research topic's direction and its underlying knowledge structure. This information is then presented in a bird-eyes view visualization format. **Chapter 5** will detail the specific application and visualization methods for this infrastructure data derived from the <Introduction> section.
- **'Insight sections' for Longitudinal insight view:** This part of dataset comprises body-text from the 'insight sections,' such as 'Conclusion' and 'Limitation' related sections of papers. Building on the bird-eyes view, this data offers a deeper summary of each paper's work and its advantages. Authors also discuss the limitations of their work and propose future research directions. This section provides readers with in-depth, summary-type assistance, exploring what problems the paper's topics can solve and what issues remain unresolved, all within a broad knowledge context. By combining 'Insight section' content from multiple articles, novice researchers gain longitudinal insight into these papers' positioning within the research topic and their interrelation-

ships. This includes elements inherited from previous research, studies addressing similar problems, or those with comparable new findings. We present this information as a Longitudinal insight view. **Chapter 6** will detail the specific application and visualization methods for this infrastructure data derived from the 'Insight sections'.

- **Whole sections for Cross-sectional insight view:** This part uses the entire body text of papers as a dataset to support cross-sectional insight. Building upon the longitudinal insight view, this survey method explores multiple viewpoints within the papers. It provides more detailed cross-sectional insights into the similarities and differences between the contents of multiple papers. These viewpoints, reached by consensus among experts in the research field, are essential investigative elements for this research topic. As the content related to these viewpoints typically spans the entire text, we use body text from various parts and their corresponding sections to refine the approximate range of each viewpoint's coverage. **Chapter 7** will detail the specific process and application of this approach.

Chapter 5

Subsystem 1 : Fish-bone diagram - Gain the bird-eyes view

5.1 Motivation

As we mentioned in **Section 2.2** ,novice researchers often struggle to comprehend numerous academic papers and grasp the fundamental concepts needed to further break down research tasks in a new field. To solve such problems, the knowledge graph supporting research survey is gradually being developed. Existing keyword-based knowledge graphs rarely provide hierarchical and logical summaries of content from multiple papers. This shortcoming hinders researchers' ability to grasp abstract concepts in a structured, logical manner. Meanwhile, as we mentioned in **Section 2.6**, novice researchers may find it difficult to use *ChatGPT* effectively for research surveys due to their limited understanding of the research field. Without the ability to ask proficient questions that align with key concepts, obtaining desired and accurate answers from large language model **(LLM)** could be inefficient. This study aims to help novice researchers by providing a ***fish-bone*** diagram that includes causal relationships, offering an overview of the research topic. The diagram is constructed using the issue ontology from academic papers, and it offers a broad, highly generalized perspective of the research field, based on relevance and logical factors. Furthermore, we evaluate the strengths and improvable points of the ***fish-bone*** diagram derived from this study's development pattern, emphasizing its potential as a viable tool for supporting research survey. As the topmost layer of our research survey assistant interface, the ***fish-bone*** diagram plays a pivotal role in guiding the overall direction of the research survey.

5.2 Fish-bone configuration

The methodology of this subsystem is divided into two parts. First, we provide a detailed definition of Issue Ontology. Second, we introduce the

fish-bone diagram, established based on Issue Ontology, to support research survey of bird-eyes view.

5.2.1 Issue ontology in 'introduction'

In academic papers, certain key elements provide insight into the author's reasoning and thought process. These elements reveal how the author discovers problems, investigates them, and contemplates solutions. This can be regarded as the writing clues of the article. The related sentences that appear throughout the papers make up the clues, which are crucial for readers to grasp the bird-eyes view of the research topic. These critical sentences encapsulate the concept of 'issue ontology' in academic articles, as defined in **Section 3.1.2**. Issues embody various debates within the academic world. Some of these have been resolved, some have been identified but remain unresolved, and others still require optimization and improvement. For most academic papers, the first section typically introduces background context that leads to the research objective. In computer science, most papers follow the *IMRAD* format [86], with "Introduction" as the first section's title, which contains specific issue ontologies [100]:

- **1. Prelude issue:** The root task mentioned in the paper reflects the historical context of this paper. It is usually in the head sentences of the 'introduction' section as the source of clues.
- **2. Improvable issue:** Some shortages the author mentioned from previous work that need to be solved.
- **3. Emphasize issue:** Reflect on the author's purpose, contribution, and what they did.

5.2.2 Design of *fish-bone*

For novice researchers, merely summarizing an overview based on keywords or sentences may not clearly convey the logical structure, which could hinder knowledge management of the research topic. To address this, we adopt ***fish-bone*** diagram - causal diagrams created by Kaoru Ishikawa, are used to display the potential causes of a specific event[1]. This type of diagram aids in understanding the learning logic from multiple articles. It includes preliminary tasks, challenges encountered, and the highlighted goals within a research topic. We present the issue ontology as a ***fish-bone*** diagram - a new type of knowledge graph for establishing the logical structure of issues in academic papers.

[1] https://en.wikipedia.org/wiki/Ishikawa_diagram

Table 5.1: Fish-bone diagram configuration

Part	**Expression**	**Function**
Head	Topic name	Present research topic.
Joint	Tasks in Topic	Represent the 'effect' in a ***fish-bone*** diagram. A research topic comprises multiple 'effects' (tasks).
Backbone	Issue Ontology in corresponding task	Express the label of 'emphasize issue' and 'improvable issue' presented in each task.
Fine-bones	Elements in Issue Ontology	Summarize the in the clustered sentences of issue ontology
Child-bone	Logic-link of issue summarization	Express the logical chain composed of 'emphasize issue' → 'improvable issue'.

Table 5.2: API and model in implementation process

Phase	**API**	**Method**	**Pre-training**
Data processing	*SpaCy*	*Sentence Segmentation*	*en_core_sci_lg*
Issue ontology classification	*Sklearn*	*Sentence embedding, SVM, Grid search*	*scibert_scivocab_cased*
Issue ontology clustering	*Sklearn*	*Sentence embedding, K-means*	*scibert_scivocab_cased*
Task name generation	*Chatgpt*	*prompt engineering*	*gpt-4o-mini*
Issue sentences summarization	*Chatgpt*	*prompt engineering*	*gpt-4o-mini*
Fish-bone diagram visualization	*Pyvis*	-	-

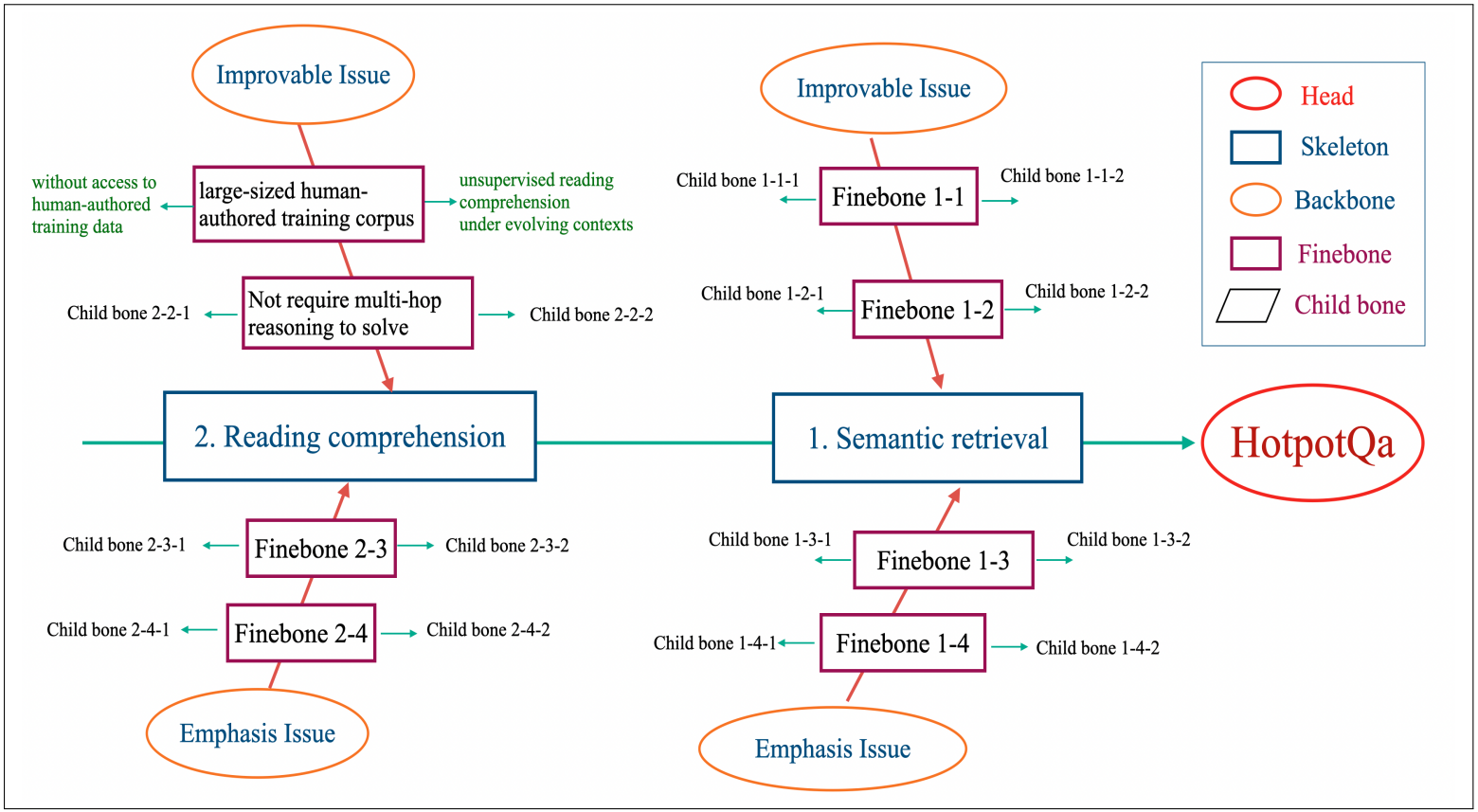


Figure 5.1: Fish-bone diagram of bird-eyes view survey

We set the number of foundational articles for composing the ***fish-bone*** diagram at approximately 500, providing directions for 5-10 research tasks. When the number of papers surpasses 1,000, the base number of decomposable research tasks becomes unwieldy. If we were to provide more than 10 research tasks in such cases, the overall summary content of the ***fish-bone*** diagram would grow exponentially. This would result in a structure so complex that novice researchers would struggle to comprehend it. If providing only 5-10 tasks when dealing with an excessive number of papers would lead to an overabundance of papers within each task. This would result in generated summaries that fail to comprehensively cover the corresponding research tasks.

5.2.2.1 Feature & Element of *fish-bone*

Fish-bone diagrams contain three key elements: effect, factor, and cause. By integrating these elements with issue ontology, the design of the ***fish-bone*** diagram can be expanded to include the feature of bird-eyes view, form the conceptual design of it as shown in **Figure 5.1** and the structure summarized in **Table 5.1**.

Using the research topic of '*HotpotQA*' [99] as an example, it includes tasks like reading comprehension, semantic retrieval, and question answering. These tasks form the 'effect' of the ***fish-bone*** as joints. Each task encompasses 'improvable issue,' which represent the persisting problems in the task. Meanwhile, 'emphasize issue' showcases the previously employed measures to tackle the 'improvable issue.' The 'improbable issue' and 'emphasize issue' constitute the backbone. Inside the backbone, the fine bone consists of sub-classes of its corresponding issue ontology, signifying the 'factor' in the ***fish-bone***. For example, in the task of reading comprehension, there are some 'improvable issues' such as 'large-sized training corpus' and 'Do not require multi-hop reasoning to solve.' One method is called 'unsupervised reading comprehension' to solve this improvable issue of 'large-sized training corpus'. At this time, we call the 'emphasize issue' corresponding to this measure as child-bone. The child-bone and fine-bone with an objective logical relationship are connected by a cause-and-effect chain to form the 'cause' in the ***fish-bone***. This provides researchers with an analysis path from task → Issue ontology → Issue logical chain, which helps researchers to capture bird-eyes view more quickly and establish connections between knowledge.

5.3 Implementation

This section describes the process of developing a ***fish-bone*** diagram that provides a bird-eyes view of the '*HotpotQA*' research topic, based on the infrastructure data. The overview of implementation is shown in **Figure 5.2**. **Table 5.2** displays the APIs and models used at each stage of development.

5.3.1 Data processing

The first step in constructing the bird-eyes view survey dataset is choosing a topic and filtering the relevant sub-dataset from the infrastructure data processed in **Chapter 4**. We iterate through each paper in the infrastructure data, pulling out those that mention '*HotpotQA*' for our sub-set. This sub-set comprises the full-text content and annotation-info of the papers linked with '*HotpotQA*.' Next, we segment the text in 'introduction' section into sentences using *'en_core_sci_lg'*model from *spaCy*[2] [101] - a comprehensive pipeline for biomedical data, including a 785k vocabulary and 600k word vectors. It provides impressive segmentation accuracy. Any complex segmentation patterns that failed were manually corrected.

5.3.2 Issue ontology classification

5.3.2.1 Human annotation

To create a machine learning dataset, experts annotated the sentence-implied issue ontology types based on the definitions provided in **Section 5.2.1**. We strictly followed the rules detailed below during this annotation process.

(1) In a few cases where two types of issue ontologies appear in one sentence, we manually split the sentence to ensure that each sentence carries only one type of issue ontology.

(2) If the author hypothesizes about a topic, it is also considered a contribution. Therefore, marked as an 'Emphasize issue.' Also, assessing the issue of a sentence in isolation is challenging, as the broader context of the targeted paper influences the annotation process.

(3) For sentences that do not fit our established issue ontology, we classify them as 'others' in our machine learning configuration. These sentences might include:

- **1.** Explanations of reasons.

[2] https://allenai.github.io/scispacy/

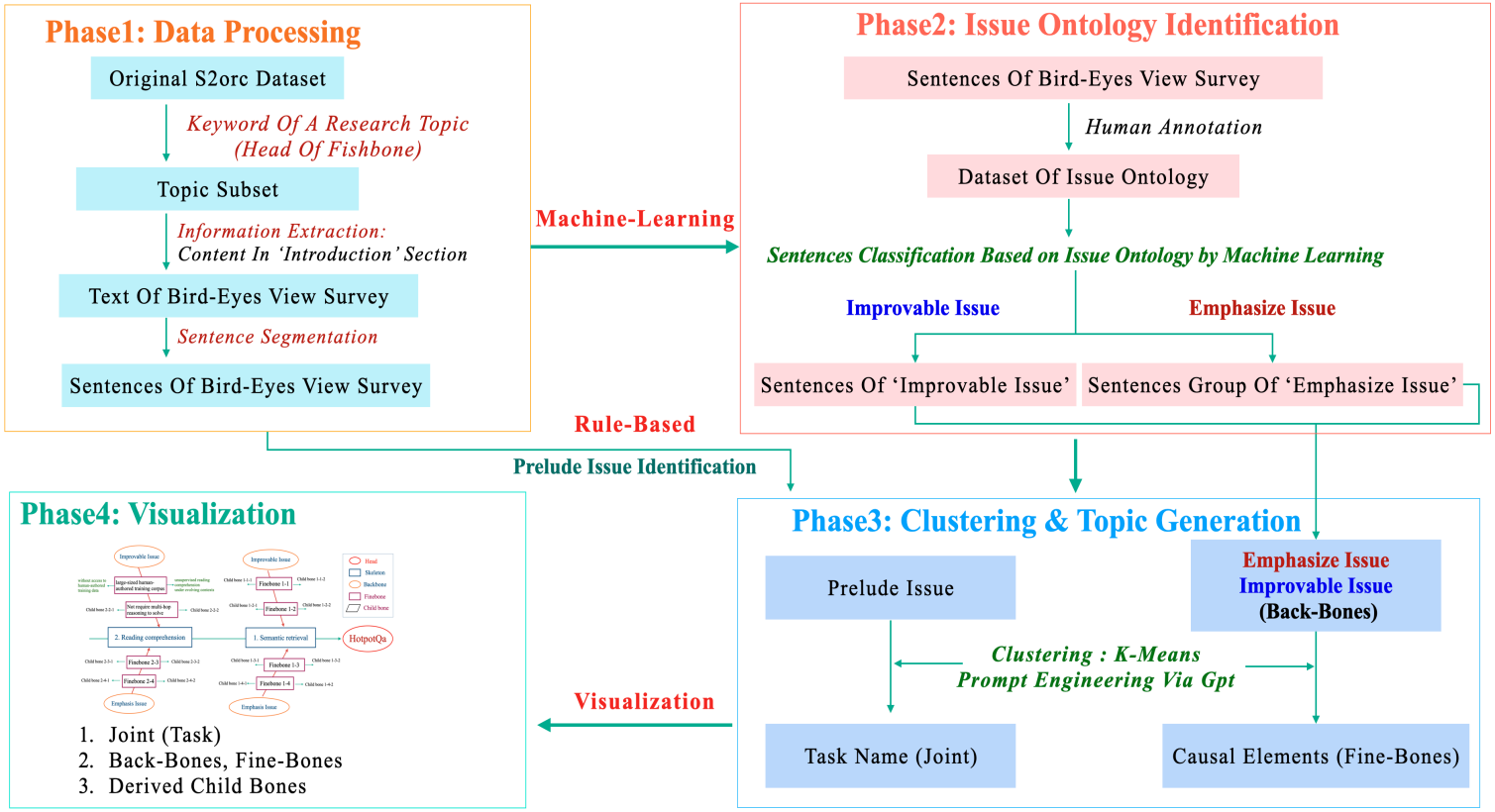


Figure 5.2: Implementation procedure

- **2.** Meaning of sentence do not match any of the three types of issue ontology.
- **3.** Issues have been addressed in previous research mentioned.
- **4.** Experimental results or performance achievements.

We found that the prelude issue typically emerges within the first 2-3 sentences of the 'introduction' section. Furthermore, the initial two sentences in most articles are usually sufficient to indicate the article's background task. Consequently, during the subsequent task clustering stage, we select the first two sentences from 'introduction' as the prelude issue for input.

5.3.2.2 Result of issue ontology classification

We utilize the *scibert_scivocab_cased* [109] pre-training model in *Sentence BERT* [102] for sentence vectorization. This model was trained using resources like Wikipedia and BookCorpus. We chose support vector machine ***(SVM)*** for classifying the vectorized sentences, as it demonstrates robust generalization performance. We also employ the grid search method [103] to identify the best parameters to apply. The optimized parameters are:

c = 10, gamma = 0.001, kernel : 'sigmoid'

The detail of dataset is shown in the **Table 5.3**. The classification results are shown in the **Table 5.4**, and the total classification accuracy has reached 78%, which proves the effectiveness of the small-scale training data. However, since sentences with 'emphasizing issue' constitute a large portion of the collected articles, the imbalance in dataset labels may influence the classification results of 'improvable issue.'

Table 5.3: Detail of issue ontology dataset of bird-eyes view survey

	Train	**Test**	**Total**
Emphasize issue	239	77	316
Improvable issue	139	48	187
Others	214	73	287
Total	592	198	790

5.3.3 Fish-bone diagram

We configure the ***fish-bone*** diagram to approach **Figure 5.1** following the description in **Table 5.1**.

Table 5.4: Classification result of issue ontology

	Precision	**Recall**	**F1-score**
Emphasize issue	0.92	0.76	0.84
Improvable issue	0.65	0.79	0.71
Others	0.73	0.80	0.76

5.3.3.1 Joint: Prelude issue clustering and task name generation

In this section, we introduce the process of using prelude issue ontology content to construct the joint (effect) of ***fish-bone***. We start by selecting the first two sentences from the 'introduction' section of each article to serve as the prelude issue. For sentence vector embedding, we also use the *scibert_scivocab_cased* model from *Sentence BERT*. These sentence vectors are then clustered using the k-means [104] method. Finally, we generate the task name for each category via simple prompt engineering by the description 'Find a theme for the following text, and the generated theme is limited to within 5 words.' These tasks form the joint (effect) component of the ***fish-bone***.

5.3.3.2 Back-bone and Fine-bone

We extend `n` branches from the task's joint to identify factors impacting each task. These branches represent the 'Emphasize' and 'Improvable' issues associated with the task. Each branch is created using the results of the set of *Sentence BERT* + issue cluster by k-means, and prompt engineering description 'Your task is to find n themes for the following text, Limit each theme to 5 words' , collectively referred to as 'fine-bones'. Here n represents the number of fine-bone to be generated. These 'fine-bones', composed of issue ontologies, significantly influence tasks, which is why we refer to them as contributing factors.

5.3.3.3 Child-bone

We build the child-bone, the most basic unit of the ***fish-bone*** design, derived from the fine-bone. This is done using the 'Improvable issue' $\longleftrightarrow$ 'Emphasize issue' logic chain within the research task. These logic connections diverge and radiate based on a cluster of real connections of the issue ontology. Specifically, we link the 'emphasize issue' that appears in the article corresponding to each fine-bone to the 'improvable issue' that appears in the same article, and then use the prompt engineering to generate

a brief summary of the sentence groups of 'improvable issue', completing the connection between the issue summaries. The process is the same when shifting from 'improvable issue' to 'emphasized issue.'

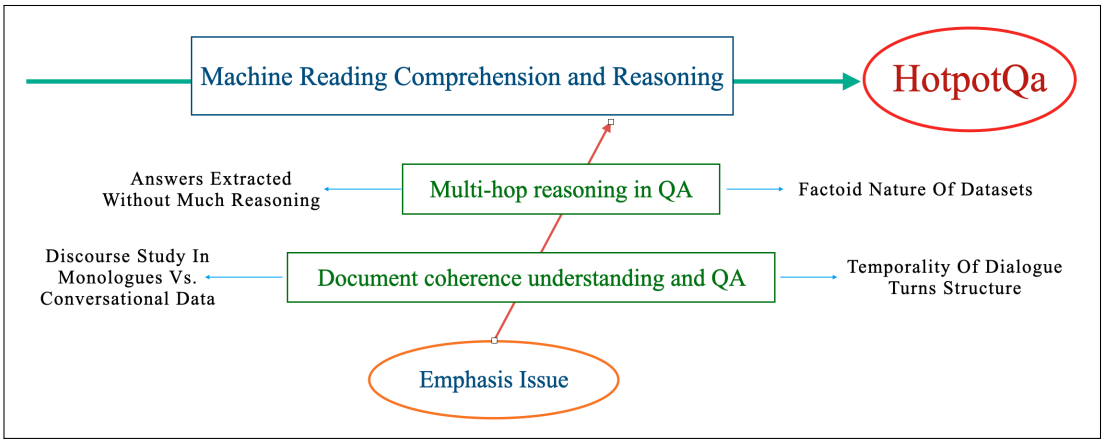


Figure 5.3: A part of ***fish-bone*** diagram

5.3.3.4 Visualization

We adopt a visualization tool, *pyvis* [105], to generate the ***fish-bone*** diagram (refer to **Figure 5.1**). This diagram offers a bird-eyes perspective and causal logic of the research topic, per the configuration outlined in **Table 5.1**.

5.4 Evaluation of fish-bone

This section evaluates and analyzes case studies of fish-bone diagrams generated from *HotpotQA* topics as described in **Section 5.3.3**

5.4.1 Sampling evaluation

To assess whether the branches in the fish-bone diagram provide factually accurate and logically structured routes (from topic → task → issue ontology), we randomly selected 20 branches and evaluated them using three criteria:

- **Consistency:** Whether the content of routes is supported by factual evidence from the Introduction sections of the *HotpotQA* source papers.
- **Reasonableness:** Whether the content of routes adhere to the logical structure of topic → task → Emphasized issue/Improvable issue → Improvable issue/Emphasized issue.

Table 5.5: Evaluation of ***fish-bone*** for 20 randomly branches

	Result (Right/Total)	**Case of defect**	**Reason of Defect**
Consistency	18/20 (90%)	Few sentences evidently do not originate from the 'introduction' section.	Some papers do not follow the *IMRAD* format [86].
Reasonableness	16/20 (80%)	Errors in issue ontology classification.	Content classified as 'Neutral' was not excluded.
Comprehensible	15/20 (75%)	Contained excessive information.	Lack of few-shot in prompt engineering.

- **Comprehensible:** Whether the content effectively highlights the main topic across multiple papers in a clear, concise way that is easy to understand.

Table 5.5 presents the evaluation results.

The results demonstrate that the fish-bone diagram effectively delivers accurate and concise knowledge in most cases, although several areas still require improvement and optimization, as detailed in **Table 5.5**. Section 5.4.2 will present typical examples for case-study

5.4.2 Case study and analysis

Using the development process described in **Section 5.3**, a part of the ***fish-bone*** diagram we simulated is shown in **Figure 5.3**. From the task of comprehending and reasoning in machine learning, we can see the logic chain of **T1** $\rightarrow$ **F1** $\rightarrow$ **C11** , which raised the issue of 'Multi-hop reasoning in QA' due to 'Answers extracted without much reasoning.' Through online searches, we found the paper 'A Survey on Multi-hop Question Answering and Generation' [106] , which states that single hop QA datasets are answerable without much reasoning, hence the birth of the idea of Multi-hop reasoning. This example proves that this logic chain indeed implies a causal relationship. However, the route derived from **T1** $\rightarrow$ **F2** $\rightarrow$ **C21**, seems inexplicable. In fact, it is mentioned in 'Discourse Analysis and Its Applications' [107] that 'monologue vs. conversation' are concepts in discourse analysis, and coherence models to evaluate monologues and conversation. We have confirmed the accuracy of the information, but there is no objective cause-and-effect relationship between the **F2** and **C21**. This may be due to the following reasons:

- There are some articles in the infrastructure data that are not in *IMRAD* format[3] [86].
- The description of the engineering prompt is unclear. It does not contain hint words showing causality and corresponding example descriptions. Thus, further development is required to enhance readability and important information identification.
- The summary output of this research is limited to five words, which may be too concise to get the important information. Therefore, future exploration is needed to control the length of the generated issue summary according to the researcher's learning characteristics.

Overall, the case-study demonstrate that the ***fish-bone*** diagram, created based on issue ontology, can help researchers understand the tasks and logic of issues in their research topic to a certain degree. It also shows acceptable rationality, correctness, and readability, proving the potential for sustainable development.

5.5 Summary of subsystem : fish-bone

This subsystem focused on automatically generating bird-eyes view ***fish-bone*** diagrams for research topics to assist novice researchers. This process uses issue ontology units, logically organized and expanded to generate the diagrams. We started by collecting introductory text from academic papers related to a specific research topic, which was then segmented into sentences level. Expert researchers annotated these sentences according to the implied issue ontology type, forming the training dataset for the bird-eyes view survey. Next, we utilized rule-based and machine learning methods to categorize and extract sentences related to prelude, improvable, and emphasized issues. We summarize the prelude issues to form the tasks of the diagram using clustering and prompt engineering. The backbone and fine-bones, which illustrate the cause-and-effect relationship, were created from the summarized 'emphasized issues' and 'improvable issues' by issue ontology classification, clustering and prompt engineering. Subsequently, we generated the child-bone from the objectives logical of 'emphasize issue' $\longleftrightarrow$ 'improvable issues'. Lastly, we evaluated the readability of the generated diagram and the sustainability of the development through case analyses. For the future expansion and improvement of this study, the following points are proposed:

[3] https://en.wikipedia.org/wiki/IMRAD

(1) Expanding the issue ontology: This study focused on automatically generating a ***fish-bone*** diagram from the 'introduction' sections of academic papers. However, issues discussed in other sections, like addressed and improvable issues in the 'related work' section, or resolved and finding issues in the 'conclusion' section, are not covered. These issue ontologies can also highlight key research points, such as problems addressed across multiple articles and future research directions. Furthermore, unlike regular papers in the *IMRAD* structure, some review papers focus on summarizing issues from previous research and speculating on potential future issues. Hence, identifying these types of issue ontologies is a task for future work.

(2) Expanding the dataset: This study relies on a small subset of the infrastructure data and focuses on a single research topic. Hence, it is necessary to expand the dataset of issue ontology to other research topics to validate its generalizability and robustness.

(3) Analysis relevance of novelty: By analyzing the relevance of novelty, we can study the novelty's relevance by combining addressed, improvable, and emphasized issues. we can deeply mine similar addressed or improvable issues across multiple articles with different emphasized issues to implement originality analysis, which can provide additional insights for researchers.

(4) Assessment Method for Academic Papers Based on Issue Ontology: We find that a promising approach for future research involves evaluating papers based on our three types of issues and ranking them in terms of completeness. For instance, if a paper emphasizes its contributions without considering the limitations, its overall score will not be assessed highly. This procedure resembles a reviewer evaluating a candidate's paper to determine its quality.

(5) Adjusting the scale of diagram and summary: A large amount of text in the overview diagram may reduce readability. Therefore, controlling the information scope based on user needs is still challenging. Along with managing the diagram's scope, ensure the summary length balances readability and provides enough detail to enhance survey efficiency for novice researchers.

(6) Application Development for Information Retrieval: This study explores the automatic generation of ***fish-bone*** diagrams with embedded causal features from issue ontology. However, it requires a more

sophisticated User Interface **(UI)** design for information retrieval, according to the data connected by the ***fish-bone*** diagram. Additionally, setting up a question-answer system based on ***fish-bone*** diagram to extract deeper essential information is a promising future direction.

5.6 Function of User Interface

Below, we introduce the **UI** functions. **Chapter 8** provides a detailed description of the subjective system evaluation by novice researchers who interacted with this subsystem through these functions. The representative **UI** is illustrated in the **Figure A.1**.

- **Function1 - Transfer the information selected for deeper exploration:** When users want to explore a task and its issues in depth, they can click on the task and its corresponding issue ontology. The relevant information then appears in the text box below, allowing users to easily review and confirm the details. Subsequently, when the user clicks the 'confirm selection' button, the chosen content is sent to *SS1* for generating the ***relevance tree***.
- **Function2 - Search-bar:** Users can search the ***fish-bone*** diagram using keywords. This function filters and displays nodes containing these keywords and their connected content, while hiding unmatched nodes. This feature helps users quickly pinpoint key issues they want to explore and trace the causal relationship chains stemming from these issues.

Chapter 6

Subsystem 2 : Tree-structured Knowledge Graph for Longitudinal Insight view

6.1 Motivation

As we mentioned in **Section 2.3 - 2.4**, novice researchers struggle to understand the directions within their research topic and the discovery of new research findings within a short time. One way to provide intuitive assistance to novice researchers is by offering relevant knowledge graphs (***KG***) and recommending related academic papers. However, existing navigation knowledge graphs mainly rely on keywords or meta information in the research field to guide researchers, which makes it difficult to clearly present the hierarchical relationships, such as inheritance and relevance between multiple related papers. Moreover, most recommendation systems for academic papers simply rely on high text similarity, confusing researchers as to why a particular article is recommended. They may lack the grasp of important information about the insight connection between 'Issue resolved' and 'Issue finding' that they hope to obtain. This subsystem aims to support research insight surveys for novice researchers by establishing a hierarchical tree-structured knowledge graph that reflects the inheritance insight and the relevance insight among multiple academic papers on specific research topics to address these issues. As the middle layer of our research survey assistant interface, it bridges the broad research tasks and issue ontology from the ***fish-bone*** diagram. This layer conducts in-depth exploration and refines insights to the individual paper level, providing novice researchers with specific research insights.

6.2 About longitudinal insight survey

We divide the Longitudinal insight into two types:

- **Inheritance insight:** A survey method that utilizes direct citation relationships among academic papers can be used to identify various branches for exploring inherited relationships in research directions. This study establishes an ***Inheritance tree*** to support this kind of insight. We also analyze the correlation of the academic issues within the components of the ***Inheritance tree***.
- **Relevance insight:** By analyzing the 'Issue finding' and 'Issue resolved' for each academic paper, a relevance chain of 'Issue Finding' → 'Issue Resolved' → 'Issue Finding' can be established between multiple academic papers to guide researchers in discovering the value of potential direction. This study introduces a ***relevance tree*** to facilitate this type of analysis.

6.3 Design of longitudinal insight view

In this subsystem, we will use '*HotpotQA*' as an example, which is a well-known dataset in the field of **NLP** [99]. The Ideal ***KG*** reflecting Inheritance insight and relevance insight survey of '*HotpotQA*' topic are shown in **Figure 6.1 - 6.2**. The configuration of the ***KG*** is shown in **Table 6.1**. In **Figure 6.1**, the branches:

(1) p1(Dataset) → p2(Reading Comprehension) → p4(Evaluation)

(2) p1(Dataset) → p3(Benchmark) → p7(Retrieval)

represent two research directions extended by the core paper of the '*HotpotQA* dataset'. This graph can help researchers infer the elements of research inheritance in the '*HotpotQA*' topic. For instance, `P1` introduced the '*HotpotQA* dataset', `P2` utilized this dataset for reading comprehension, and `P4` developed evaluation metrics based on reading comprehension. In **Figure 6.2**,

p1 → Network → p2 → Single-hop → p4

and the path of 'Retrieval' branch:

p1 → Retrieval → p3 → Iteratively → p7

are extensions of the 'Multi-hop' and 'Retrieval' research tasks in the '*HotpotQA*' topic. This tree structure included multiple paths can help researchers infer the correlation factors between subtasks in the '*HotpotQA*' topic. For example, in the 'Multi-hop' subtask, survey paper `P1` provides a list of specific tasks, `P2` extends it to the network establishment level, and `P3` incorporates single-hop methods.

6.4 Implementation procedure

We utilize body-text extracted from **CTBR** and metadata, bibliographic references from S2orc as infrastructure data of ***KG*** based on **Chapter 4**'s process. The overview of implementation is shown in **Figure 6.3**. The development process is divided into 4 sub-stages:

1. In the first stage, we meticulously manipulate secondary data based on the infrastructure data to construct our unique dataset tailored specifically for our insight survey.
2. In the second stage, we use the *Sentence Bert* model and manually set labels to perform three-class classification (Issue Resolved / Neutral / Issue Finding) on the insight content of each paper to extract the corresponding sentence.
3. In the third stage, we use the extracted sentences from phase 2 to analyze the inheritance and relevance chain in-depth. We select appropriate papers from the whole insight survey dataset based on certain criteria to generate tree-structured hierarchical trees.
4. In the final stage, we use the *pyvis* [105] to visualize the tree-structured ***KG*** and provide some case studies to demonstrate our findings.

6.4.1 Phase 1 : Data processing

To build the insight survey dataset, the first step is to select a topic and filter out the relevant sub-dataset from the infrastructure data. We iterate through each paper in the infrastructure data and extract those that contain the keyword '*HotpotQA*' into our sub-set. Our sub-set also includes data-citation information, full-text content, and meta-information of the papers associated with '*HotpotQA*'. Next, we extract the text from sections titled 'conclusion','discussion' and 'limitation' to identify insight content by using **CTBR** method mentioned **in Chapter 4**. We focus on paragraphs with section titles containing 'conclusion', 'discussion' and 'limitation' as they address the problems discussed in the research paper and highlight any

Table 6.1: Knowledge graph configuration

Property	**Content**	**Function**
Node_label	Paper title	Visually Presenting the theme of a Research Paper
Node_Title	1. Keywords of paper 2. 'Issue Resolved' in this paper 3. 'Issue Finding' in this paper	Visually display the key information of research paper, as a summary.
Edge_direction	1. ***Inheritance tree*** : Cite Source –>Cite Destination 2. ***Relevance tree*** : Issue Finding –>Issue Resolved	1. ***Inheritance tree*** : Reflects the citation direction 2. ***Relevance tree*** : Direction of Relevance chain
Edge_label	Co-occurring vocabulary	Helps user understanding of the relevance between the two papers.
Edge_Title	The value of relevance chain between papers	Make it convenient for users to explore potential direction

Table 6.2: Insight dataset processing (*HotpotQA*)

Annotation in S2orc	**Description**	**Way of Processing**
title	Paper title	Use the offset annotation of the paper title to extract the title text.
bibentry	Link cited corpusid	Find the corpusid of a paper in *S2orc* that matches the cited paper's corpusid in the '*HotpotQA*' subset.

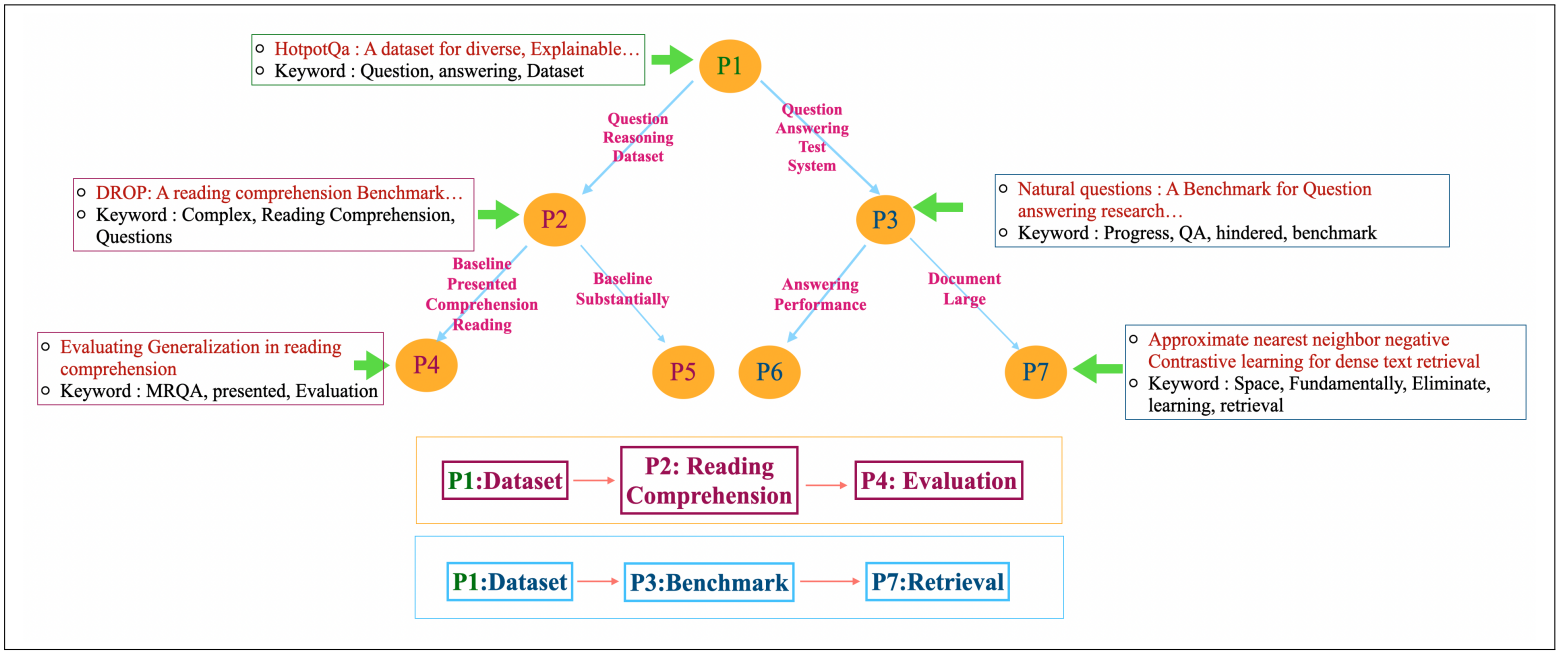


Figure 6.1: ***Inheritance tree*** plan(‘*HotpotQA* topic’)

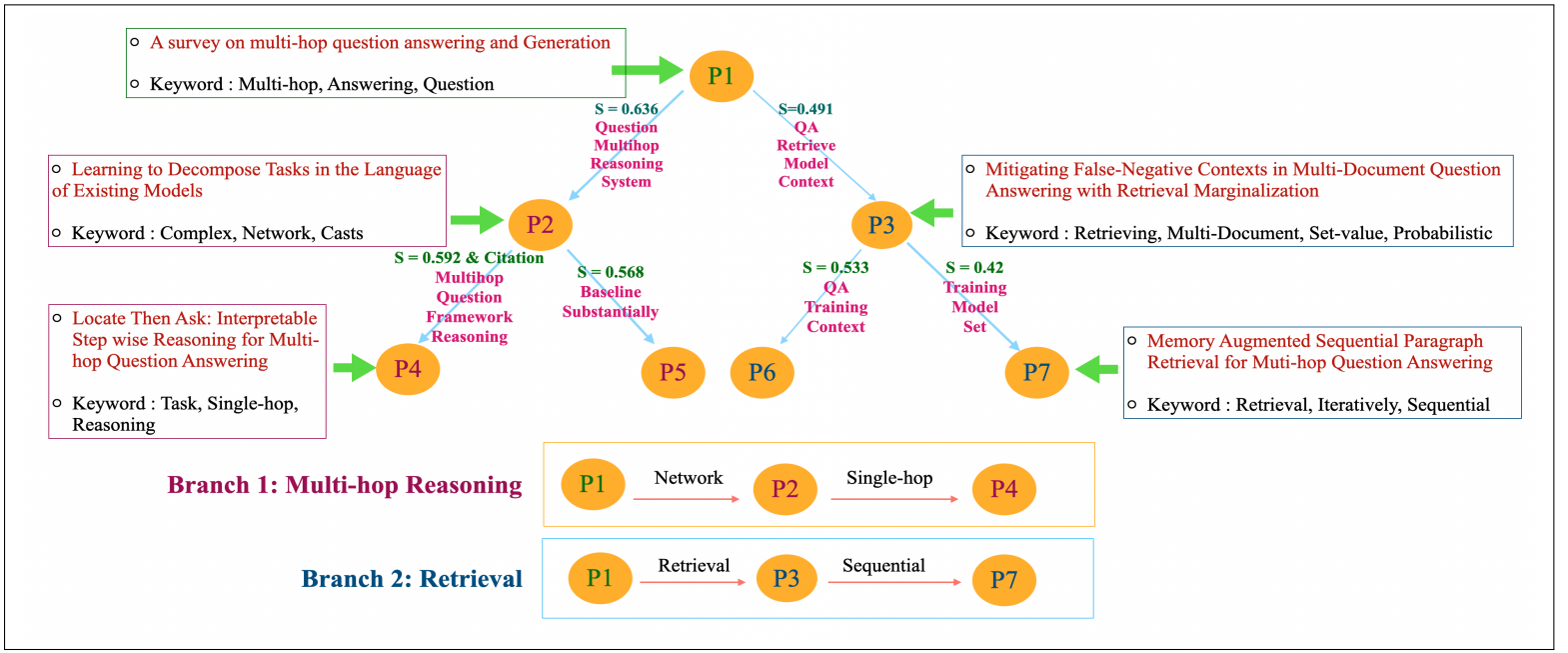


Figure 6.2: ***Relevance tree*** plan('*HotpotQA* topic')

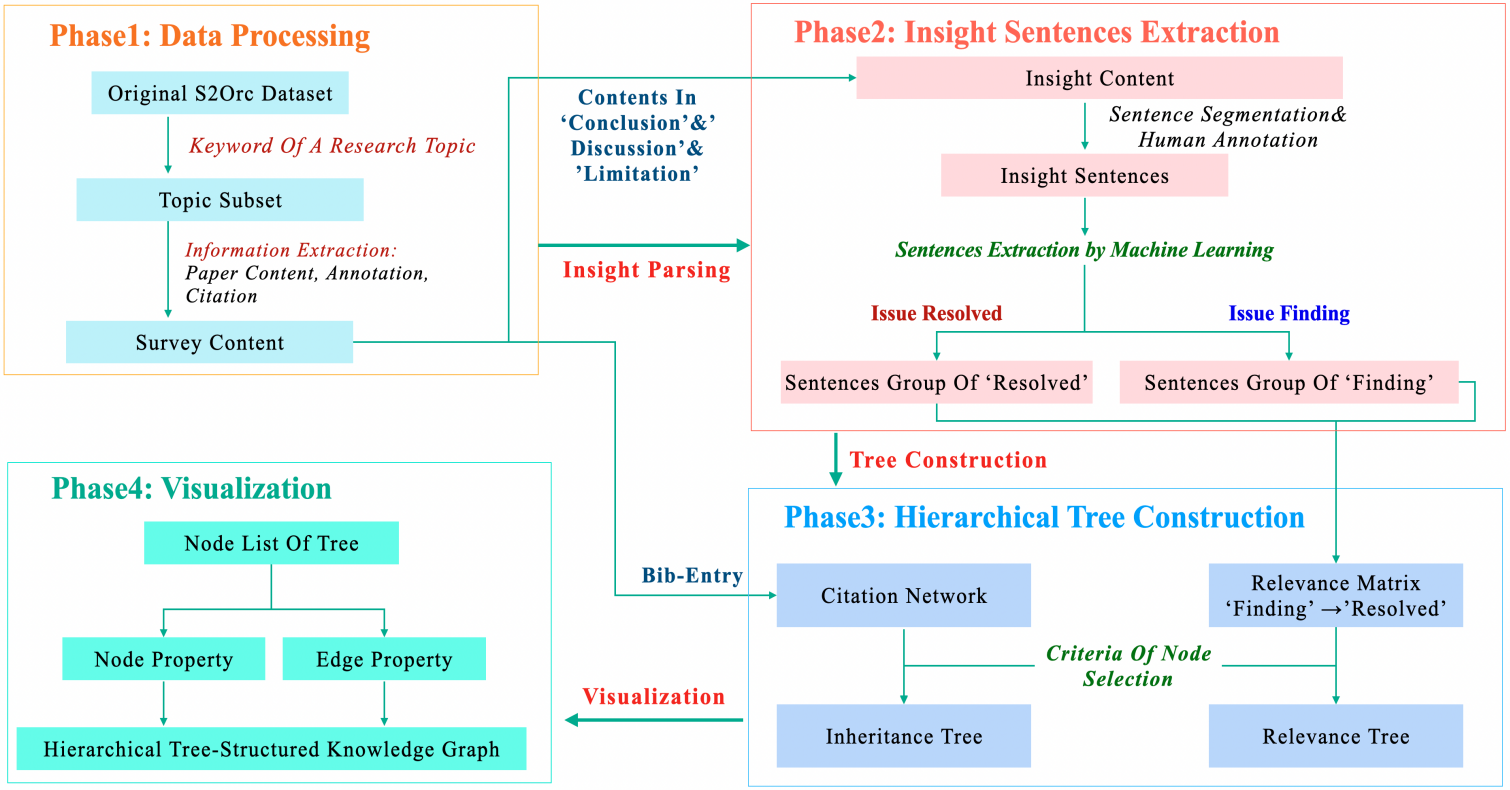


Figure 6.3: Implementation procedure

remaining challenges or limitations. Additionally, we extract relevant citation relationships from the *S2orc* dataset to create a global citation network based on the annotations of 'bib-entry'. The details of insight survey content as described in **Table 6.2**.

6.4.2 Phase 2 : Insight Sentence Extraction

In this section, we discuss the process of the insight survey dataset. First, we divide the text in the 'insight-content' into sentences. We annotate each sentence with the label 'Issue Resolved','Neutral', or 'Issue Finding' corresponding to the viewpoints the sentence expresses. Next, we use the *Sentence Bert* [102] to vectorize each sentence and adopt Support Vector Machines (***SVM***) classifier to distinguish the corresponding label for each sentence. Finally, sentences with the same 'Issue Resolved' and 'Issue Finding' labels in each article form the 'insight sentence' of that article.

Sentence Segmentation: *Spacy* is an open-source natural language processing library for Python that offers an API to access its machine learning trained methods and properties [108] [101]. This work uses the pre-trained model in *Spacy* to implement sentence segmentation. *Spacy* provides a pre-trained English library called '*en_core_sci_lg*' which includes a default sentence segmenter[1][2]. Any complex segmentation patterns that failed were manually fixed.

Human Annotation: For the segmented sentences mentioned above, experts determine the viewpoint of each sentence based on its meaning. The viewpoints include 'Issue Resolved', 'Neutral', or 'Issue Finding'. Sentences that thoroughly analyze research methods without explicitly highlighting contributions will be considered 'Neutral'. Sentences discussing potential future works in a particular field without specificity will also fall into the 'Neutral' category. 'Issue Resolved' sentences need at least a combination of contribution and experimental outcomes, regardless of their positive or negative outcomes. Sentences that ambiguously hint at trends and recommendations may be classified as 'Issue Finding'. When categorizing these three viewpoints, the references, tables, and figures of the sentences are unchanged. **Table 6.3** shows specific definitions, distinguishing criteria, and examples of these three labels. The dataset, which consists of insight sentences and labels, has been published on [3]

[1] https://allenai.github.io/scispacy/

[2] https://www.tutorialspoint.com/perform-sentence-segmentation-using-python-spacy

[3] https://www.kaggle.com/datasets/dannyleeakira/dataset-for-academic-novelty-insight-survey

Table 6.3: Human annotation strategy

Sentence	**Description**	**Label(Issue)**
In this paper, we present textbrewer, a flexible pytorch-based distillation toolkit for nlp research and applications.	The target of this paper	Resolved
Finally, we show that our model facilitates interpretability by learning an explicit hierarchy of tasks based on the skills they require.	Things achieved in this paper	Resolved
For the fourth setting, we pre-train the model for 2 epochs on the fever dataset, followed by 4 epochs on liar-plus, the fine-tune on politihop for 4 epochs.	More detail of the method they did	Neutral
We have conducted a series of experiments.	Not related to the insight viewpoints	Neutral
Future directions of our work may include using git in downstream nlp applications where the graph inductive bias is necessary and dataset is scarce.	Future work of this paper	Finding
More causal approaches such as amnesiac probing, which directly intervene in the underlying model's representations, may better distinguish between these cases.	Limitation of this paper	Finding

Training, Sentence Extraction and evaluation: We adopt *'scibert_scivocab_uncased'* pre-training model[4] [109], which was trained comprising 1.14M full-papers and 3.1B tokens, was sourced from Semantic Scholar, for sentence vectorization. *scibert_scivocab_uncased* exhibits adaptability to both the corpus and domain, making it suitable for our training data. We selected ***SVM*** for classifying the vectorized sentences due to their strong generalization performance. We also used 1500 labeled sentences for training and validation data. The training and test data details are shown in **Table 6.4**. To obtain the optimal ***SVM*** parameters, we use the grid search [103] method and find the best parameters to apply to the test. The optimized parameters are:

$$\text{c} = 10, \text{gamma} = 0.001, \text{kernel} : \text{'poly'}, \text{degree} = 3$$

The classification accuracy evaluation is presented in **Table 6.5**. The classification result shows that the 'Issue Resolved' class has a higher F1 Score, as the larger amount of data might influence it. The 'Neutral' and 'Issue Finding' classes have lower F1 scores, indicating challenges in achieving both high precision and recall. Based on the results, we extracted insight sentences of the 'Issue Resolved' and 'Issue Finding' by combining the sentences with corresponding labels within each article.

Table 6.4: Detail of issue-status dataset

	Train	**Test**
Issue Resolved	532	165
Neutral	334	121
Issue Finding	259	89
Total	1125	375

Table 6.5: Classification result of issue status

	Precision	**Recall**	**F1-score**
Issue Resolved	0.90	0.85	0.88
Neutral	0.62	0.73	0.67
Issue Finding	0.75	0.71	0.73

[4] https://huggingface.co/allenai/scibert_scivocab_uncased

6.4.3 Phase 3 - 4 : Hierarchical Tree Construction & Visualization

In this section, we utilize the sentences extracted in phase 2 and the citation information obtained in phase 1 to comprehensively extract the insight characteristics of the papers. We then employ two strategies to select specific papers to construct a tree-structured network.

6.4.3.1 Similarity calculation:

To create the ***relevance tree***, we use the classification results from Phase 2 to determine the elements of the relevance chain ('Issue finding' → 'Issue Resolved'). We then use embedded insight sentences in sentence transformers [102] with '*scibert_scivocab_uncased*' model to calculate the cosine similarity between insight sentences labeled as 'Issue Finding' and those labeled as 'Issue Resolved'. Next, we iterate through all the insight sentences in the papers and calculate the relevance chain to generate the relevance matrix. Based on the values in the relevance matrix, we select papers to construct the ***relevance tree*** in the following section.

6.4.3.2 Tree-structured *KG* Construction:

We extract specific nodes from the insight survey dataset and establish a hierarchical tree-structured according to the following rules, Where `N` represents the maximum number of root papers (The number of trees in a ***KG***), `n` is the root sequence of the selected paper, `M` represents the maximum number of leaves, `m` is the leaf sequence of the extracted paper, `T` represents the maximum depth of the tree, and `t` is the current depth of the tree.

- ***Inheritance tree*:**
  Step 1 - Root node determine: Sort all the papers in descending order based on the number of other papers that have cited them. Select the top `N` papers with the highest citation counts as the root. This operation ensures that the root node has a high level of inheritability throughout the paper library.
  Step 2 - Leaf selection: From the candidate group of papers that cite the root node paper, select the top `m` papers with the highest citation counts as the leaf nodes of root `n`.
  Step 3 - Parent node update: Set each leaf node as a new root node and repeat Step 2. If a root node is not cited (unable to generate a leaf), the branch is terminated.

- ***Relevance tree*:**
  Step 1 - Root node determine: Calculate the average similarity scores for each paper in the relevance matrix. Select the top `N` papers with the highest similarity score as the root nodes. This operation ensures that the root node has a high correlation index throughout the paper library.
  Step 2 - Leaf selection: From the candidate group of papers that have relevance chain with the root node paper. Select the top `m` papers with the highest similarity score as the leaf nodes of root `n`.
  Step 3 - Parent node update: Set each leaf node as a new parent node and repeat Step 2. If a root node does not have a relevance chain (unable to generate a leaf), the branch is terminated.

- **Common Rules:**
  Rule 1: The selected nodes cannot be selected again.
  Rule 2: Repeat Steps 1-3 until reaching the upper limit of `N` or the maximum depth limit value `T`. The network is then generated.

To provide researchers with a visual understanding (**Figure 6.1 - Figure 6.2**) of the guidance on research directions provided by these selected papers, we use *pyvis* to visualize ***KG*** based on the configuration of **Table 6.1**. We limit the number of foundational articles for constructing the ***KG*** to approximately 10-50 depended on user's option. Exceeding this range would result in an information overload within the view, potentially compromising its readability.

In addition to the co-occurring vocabulary of insight sentences between papers, we also extract keywords specific to each paper to capture its unique concept. When selecting keywords for each paper in the ***KG***, we combine the 'insight content' from all papers in the subtree to create the global text. The 'insight content' of a specific article within the subtree is used as the partial text. *Tf-idf* [110] [111] method used to extract partial text keywords from the global text composed of multiple paper-texts in a tree, and then display them as paper keywords in the node titles and a part of the representative sample branch included insight sentences is shown in **Figure 6.4 - Figure 6.5**. To prevent the occurrence of the same concept being expressed using different keywords in different articles, we adopt the *sci-vocab-uncased* pre-trained model to encode word vectors and perform cosine-similarity analysis to identify words with similar meanings [112]. Specifically, when considering a word as a keyword for a target article, we apply a condition: if this word has a threshold of **cosine-similarity < 0.6** with keywords from other articles, we

determine that it can serve as a unique keyword to illustrate the difference between the corresponding paper and other papers in the subtree.

6.4.4 Visualization of insight summary

For novice researchers, reading multiple articles with extractive summarization still presents the challenge of information overload. To address this, we adopt *prompt-engineering* via *gpt-4o-mini* **LLM** model to further condense summaries while preserving key information from the issue ontology. Our prompts incorporate the issue ontology description, with the temperature set to **0** to ensure fidelity to the original text [113]. We crafts specific prompts for 'Issue resolved' and 'Issue finding' as shown below. The resulting concise summaries, limited to `n` words, are then embedded into the ***KG*** for novice researchers to read. We set `n` as **30** in the evaluation part of **Chapter 8**.

```
prompt = f""" Your task is to extract relevant information
    from text of academic paper to make a brief summary in a
    consistent style.

     The summary should highlight [the work this paper done or
     the target this paper achieved]. Try to decrease the
    usage of adjectives and adverbs for a more concise summary
     less than {n} words.

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

Listing 6.1: Prompt - Summary of 'issue resolved'

```
prompt = f""" Your task is to extract relevant information
    from text of academic paper to make a brief summary in a
    consistent style.

     The summary should highlight [the any remaining issues
    that require further attention]. Try to decrease the usage
     of adjectives and adverbs for a more concise summary. The
     length of summary should be limited in {n} words.

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

Listing 6.2: Prompt - Summary of 'issue finding'

6.4.5 Case Study & Analysis

The internal information of sample branches in **Figure 6.1 - Figure 6.2** are shown in **Figure 6.4 - Figure 6.5**. The ***inheritance tree*** can

appropriately provide key points of clues to the direction of the research survey. However, deeply digging into the reasons for the citation is also a challenging issue. As described in **Figure 6.4**:

The left branch `P2` → `P4` shows that `P2` expects to expand in the direction of comprehensive analysis, while `P4` extends this direction to the reading comprehension task.

The right branch `P2` → `P5` shows that `P2` expects to extend distributed representation, while `P5` proposed a concrete span boundary representation.

We observe that the occurrence probability of co-occurring words in the 'insight sentences' is relatively low, as indicated by the edges of the ***inheritance tree***. This is because the ***inheritance tree*** is constructed based on citation extraction. Citations from previous studies are typically associated with the section of 'research background' or 'methodology' rather than the section of 'discussion' and 'conclusion'. This reflects the weakness of correlation in the insight survey. In contrast, the papers selected by the ***relevance tree*** exhibit higher occurrence probability and similarity of co-occurring words in the 'insight sentences.' As described in **Figure 6.5**:

The left branch `P1` → `P2` shows that `P1` prompts the models to perform multi-hop reasoning, while `P2` provides a text modular network that can perform multi-hop reasoning with a state-of-the-art model.

The right branch `P1` → `P3` shows that `P1` remains a challenging direction of context retrieval, while `P3` provides a new retrieving set-valued context.

Based on the above, it is evident that there are two main research trends in the task of multi-hop questions: multi-hop reasoning and retrieval. Both of them provide extensive discussions on the QA model. Thus, this type of graph offers researchers an overview before they delve into the details. Although the chain from 'Issue Finding' to 'Issue Resolved' may not be completely reasonable in content, it provides researchers with certain keywords and guidance for research directions. On the other hand, the ***inheritance tree*** serves as a broader extension based on the research tasks. However, there is a need to incorporate the similarity of relevance between papers horizontally to expand the breadth of the ***KG*** and enhance the effectiveness of the ***inheritance tree*** in supporting researchers' understanding of research evolution.

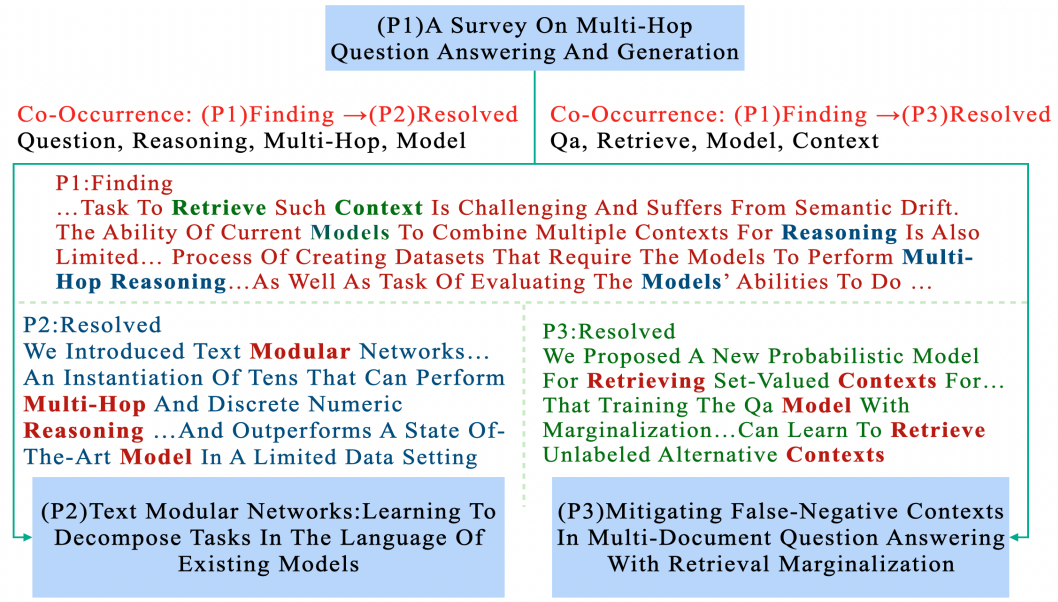


Figure 6.4: ***Inheritance tree*** Sample branch(p2,p4,p5 in **Figure 6.1**)

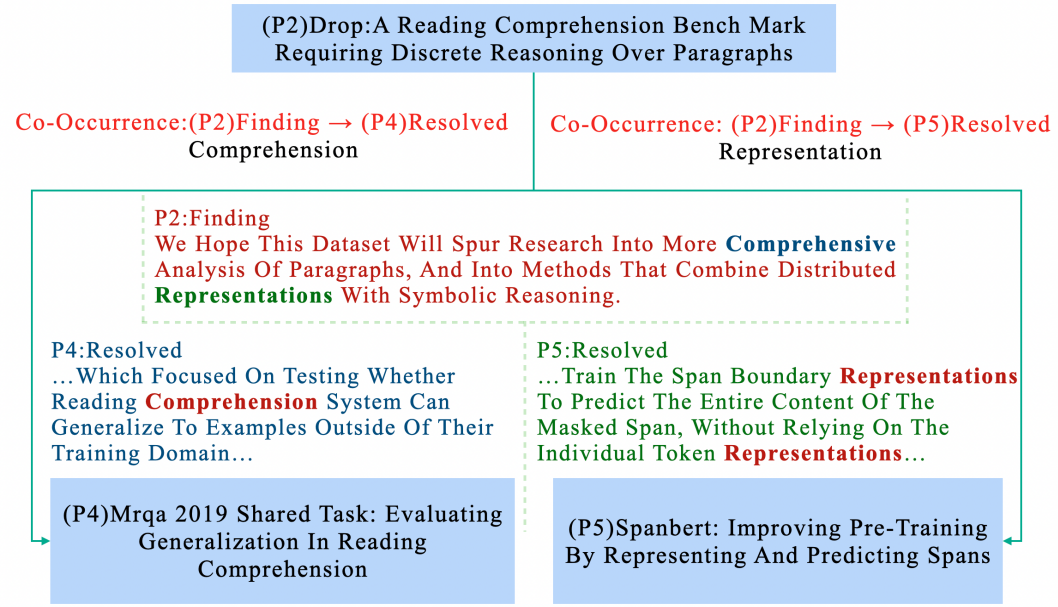


Figure 6.5: ***Relevance tree*** Sample branch(p2,p4,p5 in **Figure 6.2**)

6.5 Evaluation of relevance tree

The relevance tree developed in this chapter serves as a subsystem within the research survey assistant interface. This section evaluates and analyzes the concise summary outputs presented in **Section 6.4.4**, which are generated from relevance trees built using *HotpotQA* topics as described in **Section 6.4.3**. To evaluate whether the branches in the relevance tree diagram follow a logical structure (Finding issue → Resolved issue → Finding issue) and express their unique point, we randomly selected 20 branches with each branch containing three papers (Total 32 papers with 204 unique keywords, In some routes, multiple child nodes are derived from the same parent node), and assessed them using two criteria :

- **Correctness of unique expression:** Whether the target paper contains keywords that distinguish its unique characteristics from other papers in the branch.
- **Reasonableness of route:** Whether the route adheres to the basic logical structure of Finding issue → Resolved issue → Finding issue. Similarities between papers can be confirmed without requiring complete alignment with the issue chain logic (where Finding issue in one paper correspond directly to Resolved issue in subsequent papers). In other words, Whether research directions can be expanded by identifying similar patterns across insight summaries.

Table 6.6 presents the evaluation results. The results indicate that, in addition to improving the classification accuracy of the issue ontology, more refined prompts are needed in the relevance tree's summary generation phase to create summaries that emphasize connections between paper pairs, rather than generating isolated summaries of individual papers. Additionally, to accurately highlight unique keywords from each paper, we need to pre-identify common expressions within the field and include them in the stopword list to improve the quality of unique keyword extraction.

6.6 Summary of subsystem - Insight tree

This subsystem developed two types of hierarchical tree-structured knowledge graphs called ***inheritance tree*** and ***relevance tree*** to support insight surveys for novice researchers. Different from the previous academic ***KGs***, we define Insight survey, expanding knowledge mining in the research topic from insight perspective, promoting researchers to potentially gain insights into research directions efficiently. The process of development consists of

Table 6.6: Evaluation of relevance for 20 randomly branches

	Result	**Case of defect**	**Reason of Defect**
Correctness of unique expression	167/204 (81.9%)	Fixed expressions commonly found in papers like 'background' and 'propose' do not capture the unique aspects of the research	The stopword list was not customized to account for common words in the research field before identifying keywords by *tf-idf*.
Reasonableness of route	15/20 (70%)	Some route summaries failed to demonstrate clear connections within their relevance.	Incorrect issue ontology classification and key information being omitted during prompt engineering due to excessive focus on conciseness.

four stages: data processing, insight sentence extraction, hierarchical tree construction, and ***KG*** visualization. First, we conducted high-dimensional secondary development from the infrastructure data to create an insight survey dataset that includes citation information and insight content. Then, we extracted sentences from the insight survey dataset that express insight viewpoints on 'Issue finding' and 'Issue resolved' using machine learning techniques. These sentences were parsed and used to construct a relevance matrix. Finally, based on the citation information and relevance matrix, we generated two types of hierarchical tree structures: ***Inheritance*** and ***Relevance tree***. The generated ***KG*** demonstrates their rationality, indicating that they can provide key-information to assist researchers in gaining insights into the directions of the research topic. The knowledge graph also exhibits interpretability and potential for further development. For the future expansion and improvement of this study, the following points are proposed:

1. Incorporate Additional Text for Text Similarity and Relevance Chain Computation: Currently, this study only utilizes the content from the 'conclusion', 'discussion', and 'limitation' sections. However, the sections 'abstract' and 'introduction', specifically the part discussing previous issues, contain insight elements related to 'issue need to be solved in previous research' and the 'objective in this study'. Therefore, integrating these texts for high-dimensional training can optimize the content in the knowledge graph.
2. We objectively evaluated the classification accuracy of the 'Issue Finding' and 'Issue Resolved' viewpoints. However, a future challenge is incorporating researchers' subjective evaluations into the generated knowledge graph.
3. Insert a timeline sequence of causal chain for the research branch into

the ***KG*** to conduct comprehensive research issue analysis, including historical issues, current issues that are developing, and potential pending issues. This approach will assist researchers in better understanding the unique characteristics of the research branch.

4. This study categorized insight content and constructed ***KG*** based on it. Researchers still need to read a considerable amount of text to comprehend its specific relevance. Thus, in the subsequent step, we will leverage prompt engineering and large language models to implement abstractive summarization for insight sentences. This will provide more concise and efficient information for novice researchers.

6.7 Function of User Interface

Below, we introduce the **UI** functions. **Chapter 8** provides a detailed description of the subjective system evaluation by novice researchers who interacted with this subsystem through these functions. The representative **UI** is illustrated in the **Figure A.2**.

- **Function1 - Node summary details:** As shown in the figure, each node contains unique keywords and a summary of the corresponding paper. When users hover their cursor over a node, a summary label automatically appears. This feature allows users to quickly grasp the overview of an article.
- **Function2 - Issue summary in edge:** As shown in the table, the edge displays issue summaries for two connected nodes. It includes the Finding issue summary of the parent node `x0` and the Resolved issue summary of the child node `x1`. These 30-word summaries are generated using the prompt-engineering method described in **Section ??**. When users hover over an edge, its summary label appears automatically. This feature helps users deeply understand issue relationships across multiple papers, facilitating exploration of issue relevance.
- **Function3 - Transfer the selected paper:** This function connects with the diff-table function. For an in-depth cross-sectional comparison of papers from the ***relevance tree*** across multiple perspectives, users should double-click on two or more papers they want to explore further. After confirming their selection, users can click the 'Confirm selected paper' button. This action sends the chosen papers to the back-end system, which then generates the diff-table for cross-sectional survey.
- **Function4 - Search-bar:** Similar to the Fish-bone **UI**'s search function in **Chapter 5**. It filters and displays node titles and adjacent

nodes that match the search string. This feature helps users quickly identify relevant nodes from multiple paper titles.

Chapter 7

SubSystem 3 : A Viewpoints Embedded *Diff-table* System For Cross-sectional Insight View

7.1 Motivation

This chapter explores the final stage of the Research survey assistant: the diff-table generation method. **In Section 7.1**, we:

- Introduce the diff-table's role in the overall system
- Compare the diff-table to the bird-eyes view (**Chapter 5**) and the Longitudinal insight view (**Chapter 6**)
- Explain how the diff-table integrates with the top-down survey logic

In **Section 7.2**, we present an introduction of the diff-table subsystem, introducing its research objectives and significance.

In the **Chapter 5 - 6**, we mainly assisted novice researchers from two perspectives in conducting their research surveys more efficiently:

- *The bird-eyes view survey*, which determines the causal logic research issue [114].
- *The longitudinal insight survey*, which analyzes the relevance and inheritance among articles [59].

Both of them rely on 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 [115], allowing for cause-and-effect comparisons across multiple papers, and it enables 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's 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.

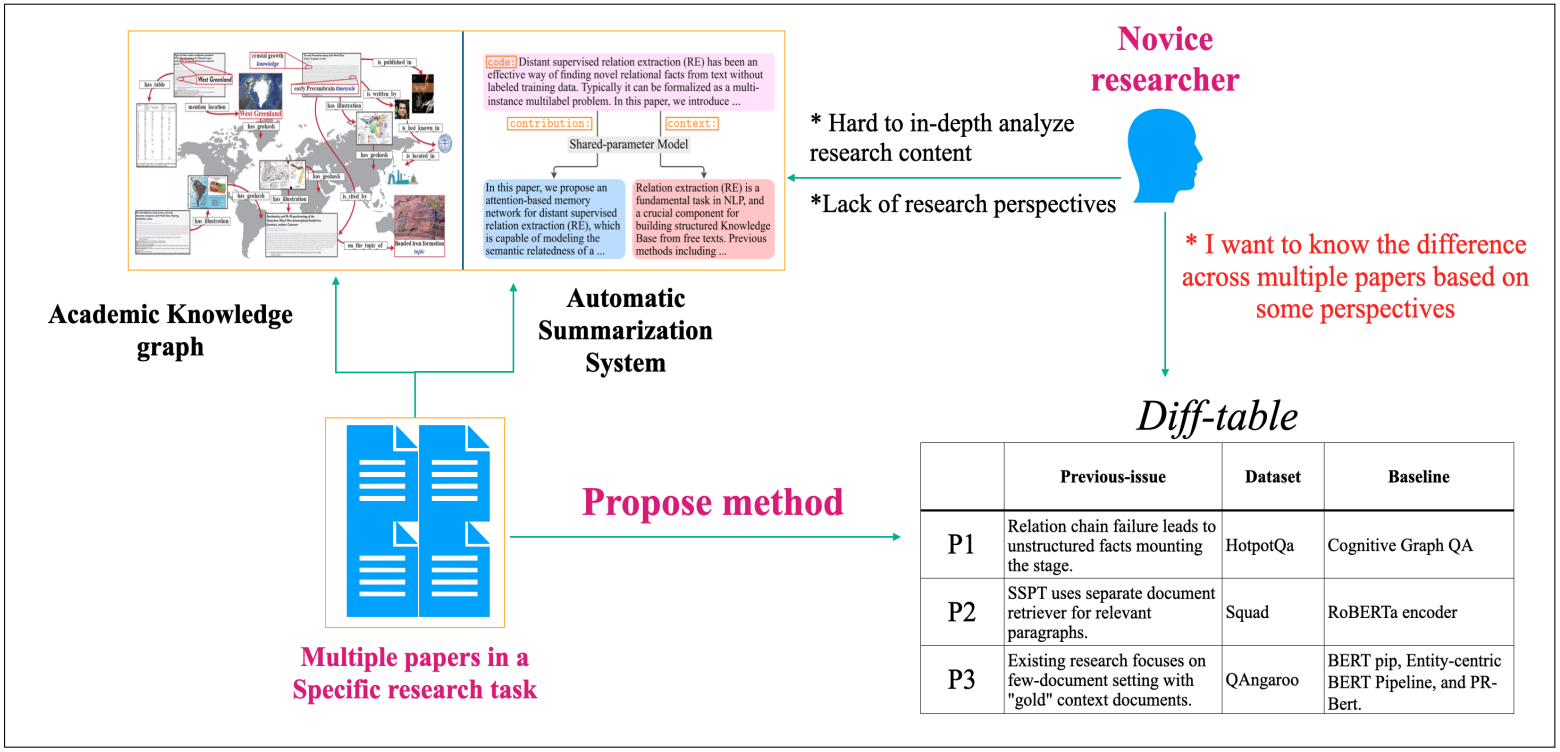
7.2 Objectives and significance of diff-table

One way to support the cross-sectional research insight survey is using prompt engineering based on *ChatGPT* to generate abstractive summarization [116, 117]. 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, *ChatGPT*'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 [118,119].

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 7.1**, 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 insights research surveys. We specially develop a dataset based on infrastructure data built from **Chapter 4** 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***.



	Previous-issue	Dataset	Baseline
P1	Relation chain failure leads to unstructured facts mounting the stage.	HotpotQa	Cognitive Graph QA
P2	SSPT uses separate document retriever for relevant paragraphs.	Squad	RoBERTa encoder
P3	Existing research focuses on few-document setting with "gold" context documents.	QAngaroo	BERT pip, Entity-centric BERT Pipeline, and PR-Bert.

Figure 7.1: The feature of diff-table, different from academic ***KG*** [120] and automatic summarization system [61]

.

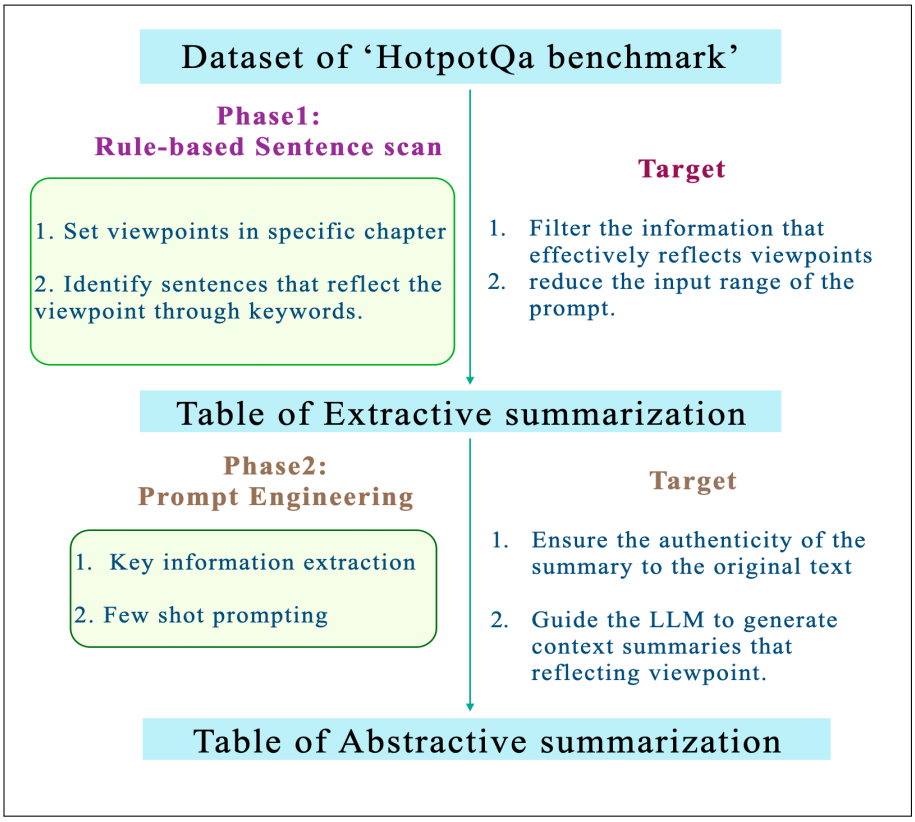


Figure 7.2: Pipeline of ***diff-table*** system development

7.3 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 7.2**. 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.

7.3.1 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 [121,122] 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 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 7.1**. 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.

7.3.2 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.

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

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

Table 7.1: 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

7.3.2.1 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 [123], incorporating an example into each prompt description to guide the model towards context imitation. **Table 7.2** presents an example of each viewpoint summary.

The sample description of prompt in the information 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 7.1**).

```
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 7.1: Prompt - Viewpoint-text Identification

7.3.2.2 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

Table 7.2: Sample summary used in few shot prompt engineering - Original text extracted from [124]

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

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 7.2: Prompt - Compression

7.4 Evaluation of summaries in *Diff-table*

This chapter evaluates the summarization generated in the ***diff-table*** from both objective and subjective perspectives. The summaries produced in **Chapters 5 – 6** for the ***fish-bone*** and ***KG*** of longitudinal insight survey, which are based on numerous articles, are challenging to measure using human-defined gold standards for quality assessment. This difficulty arises from the lack of fair indicators to subjectively evaluate the accuracy of summaries for such abstract views. For the ***Fish-bone*** and ***KG*** of longitudinal insight survey, We can only objectively evaluate the accuracy of extractive summarization through the classification results of the issue ontology (mentioned in **Section 5.3.2** and **Section 6.4.2**) to ensure that the direction of the generated views aligns with our intended purpose. In contrast, the ***diff-table*** in this chapter uses clear generation criteria for research viewpoints across multiple articles. Its more specific summary content enables the establishment of human-defined gold standards, allowing for a thorough evaluation of the ***diff-table*** summaries' quality.

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

Table 7.3: Evaluation of abstrative summrization - **Left: *BERTScore*(Average F_1) — Right: Redundancy rate**

	Our approach (Zero-shot)	**Our approach (Few-shot)**	***Scispace*** **(Baseline)**
Previous-issue	0.66 — 2.49	**0.71** — 1.48	0.60 — **1.45**
Objective	0.73 — 2.47	**0.76** — **1.33**	0.65 — 1.94
Dataset	0.66 — 0.99	**0.72** — **0.89**	0.64 — 1.46
Pre-training	0.66 — **0.35**	**0.70** — 0.59	0.59 — 1.05
Baseline	0.61 — 3.40	**0.65** — 2.35	0.63 — **2.14**
Performance	0.67 — 2.05	**0.68** — 1.84	0.67 — **0.90**
Limitation	0.59 — 0.94	0.59 — **0.48**	**0.61** — 1.51
Future-work	0.69 — 2.58	**0.75** — **1.07**	0.64 — 1.24

Table 7.4: Case study of ***diff-table*** for 'Objective' : Sample paper [125]

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.

evaluate and compare the abstractive summarization in ***diff-table***. Lastly, we subjectively evaluate of abstractive summarization in ***diff-table*** from four perspectives: Consistency, Correctness of Viewpoint ***(VP)***, Comprehensible, and Sufficient Coverage ***(SC)*** to validate the effectiveness of ***diff-table***.

7.4.1 Data-processing

This study uses data from the *HotpotQA* benchmark task [99], 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 infrastructure data processed in **Chapter 4**. 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.

7.4.2 Gold standard

To objectively and subjectively evaluate the performance of the generated summarization, we reviewed the target articles and established a gold standard, following the writing standards based on the definition of viewpoint in **Table 7.1** and the output features (summary style) in **Table 7.2**. 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.

7.4.3 Evaluation via *BERTScore*

To objectively evaluate the generated summaries in ***diff-table***, we use *BERTScore* [126] 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, such as *Scispace*[4]. 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.

[3] https://paperswithcode.com/sota/question-answering-on-hotpotqa
[4] https://typeset.io

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_uncased*[5] pre-training model, which was trained using a corpus of scientific papers, as the evaluation model for *BERTScore* [109]. 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 7.3**. 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.

7.4.4 Evaluate through human reading effectiveness

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

[5] https://huggingface.co/allenai/scibert_scivocab_uncased

not sufficient to accurately measure the effectiveness of the summary. One case study illustrates that compared to the gold standard shown in **Table 7.4**, 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 assessment of the summary's quality. To improve the shortage of evaluation via *BERTScore*, we refer to the definition of [127, 128] 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) [129]

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.

Step1:

- **-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***.

Step2: 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:

$$SC = \frac{Count_{fully\_expressed}}{Count_{GD}} \tag{7.1}$$

$Count_{fully\_expressed}$: 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.

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 7.5** 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 7.5 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 7.1** 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.

Table 7.5: Subjective Evaluation - The average score of 10 articles for each viewpoint: Correctness &Comprehensible (**C**), Sufficient Coverage(**SC**)

	Zero-shot	**Few-shot**	**Scispace**
Previous issue	C : 0.10	**C : 0.70**	C : -0.10
	SC : 0.64	**SC : 0.79**	SC : 0.22
Objective	C : 0.00	**C : 0.70**	C : 0.20
	SC : 0.79	SC : 0.76	SC : 0.63
Dataset	C : 0.00	**C : 0.40**	C : 0.20
	SC : 0.55	**SC : 0.73**	SC : 0.55
Pre-training	C : 0.00	C : 0.00	**C : 0.33**
	SC : 0.40	SC : 0.46	**SC : 0.61**
Baseline	C : 0.10	**C : 0.30**	C : - 0.30
	SC : 0.55	SC : 0.54	SC : 0.48
Performance	C : 0.20	C : 0.50	**C : 0.70**
	SC : 0.59	**SC : 0.59**	SC : 0.50
Limitation	C : - 0.20	**C : 0.00**	**C : 0.00**
	SC : 0.22	**SC : 0.22**	**SC : 0.22**
Future work	C : - 0.11	**C : 0.56**	C : 0.11
	SC : 0.52	**SC : 0.63**	SC : 0.37

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 7.6,7.7,7.8** 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.

7.5 Summary of this subsystem

We proposed a ***diff-table*** system for cross-sectional research insight survey, aimed at aiding researchers in identifying similarities and differences in research task through cross-comparison. Relying 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 our evaluation experiment, we assessed the comprehensible, minimal, and sufficient of the summaries in ***diff-table***, using both objective measures such as *BERTScore* and subjective evaluations. Importantly, the ***diff-table*** holds potential for supporting cross-sectional research surveys, providing a promising direction for future development. For the future expansion and improvement of this study, the following points are proposed:

1. Machine learning technology for Extractive summarization: This study employed keyword scanning to extract sentences that express viewpoints. However, this method may struggle to identify sentences that don't align with our established rules, such as sentence shown bellow that discussing previous issues that don't 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

Table 7.6: 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 7.7: 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 7.8: 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

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 benchmark of the pre-training model, are found in the article's table and are not included in the body-text. Therefore, it is equally important to identify and extract this kind of multi-modal information.

2. Expression of Longitudinal Knowledge Structure This study mainly focuses on the Cross-sectional Insight Survey. Based on these findings, the expression of combination with 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.

7.6 Function of User Interface

Below, we introduce the **UI** functions. **Chapter 8** provides a detailed description of the subjective system evaluation by novice researchers who interacted with this subsystem through these functions. The representative **UI** is illustrated in the **Figure A.3**.

- **Function1 - Search bar:** Highlights all cells containing user-specified keywords, facilitating easy identification of similarities across multiple papers.
- **Function2 - Pop-up window for comparison:** When a user double-clicks a cell's content, a pop-up window appears displaying that content. Users can freely drag this floating window to reposition it. This feature, combined with the search bar's highlight function, enables users to first locate multiple cells containing specific keywords, exploring their commonalities. Then, by repositioning the pop-up windows of these cells, users can easily compare the differences in content among cells with similar information.

Chapter 8

Subjective Evaluation of Research Survey Assistant Interface

Our experiment aims to evaluate the effectiveness of our survey assistant interface in helping novice researchers explore new research topics. We first measure the performance and user experience of each subsystem function independently as a baseline. Then, we analyze how well the top-down level triggers facilitate the research process compared to the independent subsystems without triggers. Through this subjective evaluation, we aim to quantify the system's impact on research survey efficiency and quality for novice researchers.

8.1 Experiment setting

8.1.1 Dataset

Using the *S2orc* dataset and the **CTBR** method for extracting sections and body-text from papers, we created datasets for two target research topics: (1) *HotpotQA* for question answering and (2) *CNN/Daily* for text summarization. The details of these datasets are presented in the **Table 8.1**.

Table 8.1: Experiment data (Topic : *HotpotQA & CNN/Daily*)

	Number of papers	**Number of sentence**
HotpotQA	489	4979
CNN/daily	659	5861

Our system generates datasets for papers on these two topics based on the description in **Chapter 4**, serving as the foundational data for visualization. The aim is to allow novice researchers to quickly grasp an overview of this vast number of papers, and to stimulate a certain degree of exploration into research directions based on this overview.

8.1.2 Experiment overview

1. Baseline Setting: Users experience individual subsystems as a baseline without **UI** operations triggers. In this scenario, users can grasp the subsystem elements to some extent, but the goals and research directions of each subsystem remain unclear. For example:

- The Fish-bone diagram's overly condensed summaries significantly limit its effectiveness. It hinders deeper exploration of research topics beyond surface-level concepts. This brevity, while intended for conciseness, unintentionally prevents novice researchers from gaining a comprehensive understanding.
- In the Relevance tree, the scope and boundaries of related papers are not clearly defined, and the criteria for including specific papers within this domain lack transparency. This ambiguity in both the domain's extent and the selection rationale for included papers can potentially lead to confusion or misinterpretation of the research landscape, especially for novice researchers attempting to navigate the field.
- In ***Diff-table***, the long paper summaries, coupled with a lack of context regarding the paper's position in the whole research topic, diminish the motivation to read.

As a result, the top-down approach in the overall survey may not be effectively implemented through independent subsystems.

2. System coherence enhancement: After enhancing system coherence(baseline improvement), users experience the entire system with **UI** operations triggering incentives. This allows users to complete the top-down logic by identifying generative factors linking the three views. For instance, Fish-bone highlights key issues and related topics within the subject area. Relevance tree then explores highly pertinent papers based on these important issues. Finally, Diff-table generates detailed summaries of these papers for comparison. Through this process, users gain a better understanding of knowledge structure, hierarchical relationships, and causal connections, which in turn motivates them to conduct further research in a clear target.

8.1.3 Criteria

We invited 11 graduate students (comprising both master's and doctoral candidates) to participate in the experiment. The table below illustrates their academic experience and research fields.

To assess whether the overall interface has guidance advantages com-

pared to independent subsystems, we have established three main aspects to evaluate novice researchers' user experience with the entire system and its individual subsystems. Participants are required to score the following aspects:

- **Functional consistency:** In the generated visualization tool, assess whether the content displayed in nodes and edges aligns with the described function, and whether provides users with effective prompts for a research overview.
- **Survey continuity:** Evaluate whether users can discern logical connections between concepts by reading this visualization tool, thereby gaining insights and directions for their research survey.
- **Comprehensibility:** Assess whether the visualization tool's presentation is concise and facilitates easy comprehension of the survey elements.

8.1.4 Experiment flow

Given the extensive information available in the target topic, we needed to limit the amount of content shown in each generated view during experiments, despite providing concise summary views. These limitations were necessary due to time constraints, as detailed in the following approaches:

1. **Fish-bone:** In the initial generation of the ***fish-bone*** diagram, we set specific parameters to provide an appropriate amount of information for novice researchers. The diagram includes 5 main tasks (represented by 5 joints), with each task containing 5 improvable issues and 5 emphasized issues (resulting in 10 'fine-bones' per joint). Additionally, each fine-bone branches into 2 child-bones.
2. **Relevance tree:** Presenting excessive paper information to novice researchers lacking background knowledge can impede effective information absorption within limited time constraints. To control information density in the generated relevance tree, we implemented three constraints in our experiments: limiting trees to two branches per node, setting a maximum depth of three levels, and restricting the total number of papers to 20.
3. **Diff-table:** To manage the ***diff-table***'s complexity and ensure novice researchers can thoroughly examine paper details in a limited time, we limit participants to selecting up to 5 papers from the ***relevance tree*** before generating their ***diff-table*** view.

Based on these experimental view parameters, the detailed setting of our

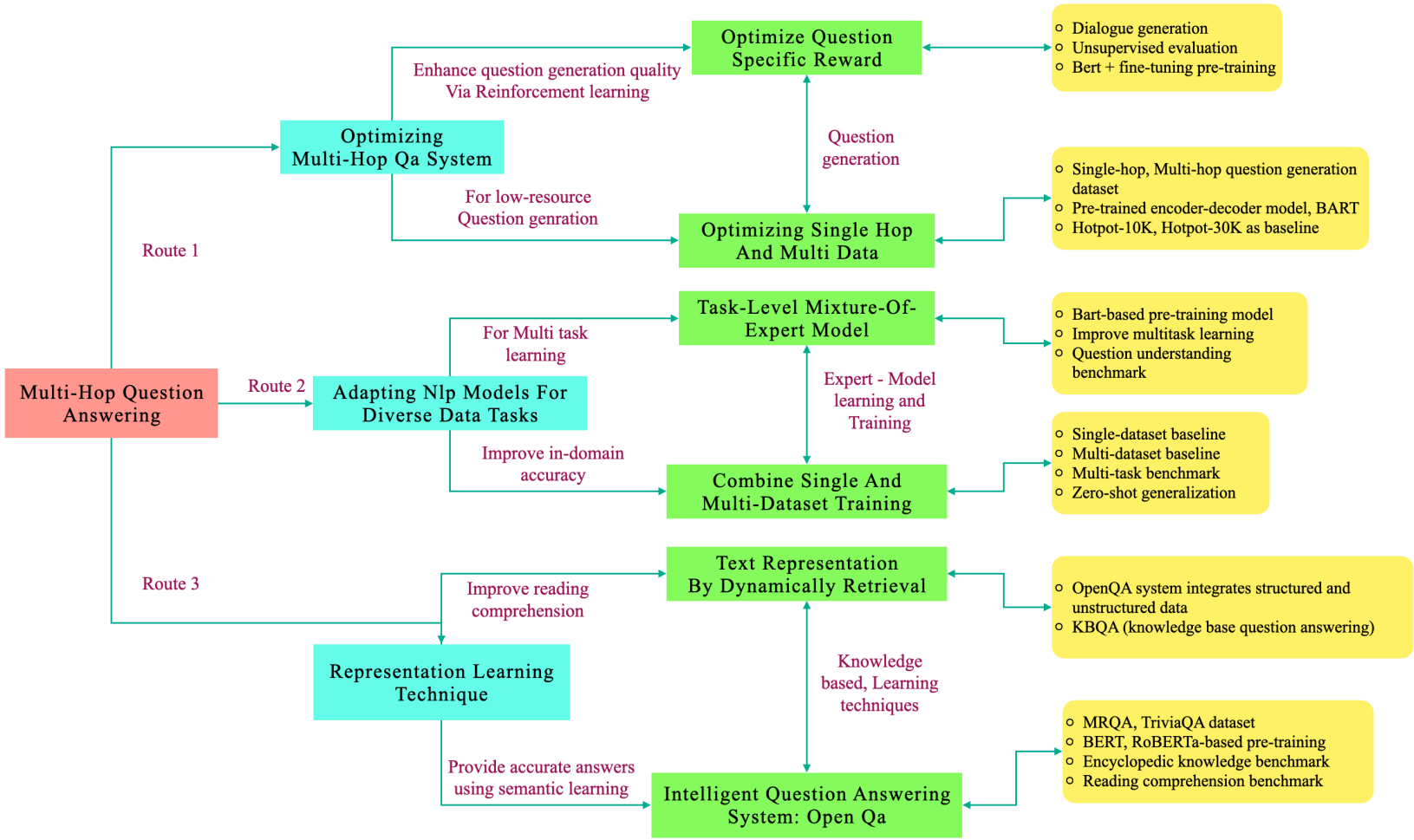


Figure 8.1: Example of top-down survey route diagram

comparative experiment are shown below:

- **Experiment1 - Hands-on subsystem (15 minutes per subsystem x 3 = 45 minutes):** Participants will engage in a hands-on experience of each subsystem, evaluating it from three aspects: functional consistency, survey continuity, and comprehensibility.
- **Experiment2 - Hands-on Research assistant interface (30 minutes):** Participants will test the overall system's functionality, which integrates three interconnected subsystems. They'll assess improvements in functional consistency, survey continuity, and comprehensibility. Afterward, they'll note the most impaction top-down survey logical elements from this interconnected experience.

We assign two topics to each group to mitigate potential bias arising from participants' prior knowledge of the research topic. This approach ensures a more balanced and fair experimental design. We divided the 11 participants into 4 groups, with Groups 1-3 having 3 participants each, and Group 4 having 2 participants. The experimental grouping for the 11 participants is outlined as follows:

- **Group 1: Experiment 1 (Topic 1) → Experiment 2 (Topic 2)** : This group begins with Topic 1 in the first experiment and then transitions to Topic 2 in the second experiment, allowing for a comparison of performance and understanding across different subject matters.
- **Group 2: Experiment 1 (Topic 2) → Experiment 2 (Topic 1)** : In contrast to Group 1, this group starts with Topic 2 and then moves to Topic 1, enabling an assessment of how prior exposure to one topic might influence performance on another.
- **Group 3: Experiment 2 (Topic 1) → Experiment 1 (Topic 2)** : This group reverses the order of experiments compared to Group 1, which helps control for any potential order effects in the experimental design.
- **Group 4: Experiment 2 (Topic 2) → Experiment 1 (Topic 1)** : This group reverses the order of experiments compared to Group 2, which helps control for any potential order effects in the experimental design.

This experimental design allows for a comprehensive analysis of both topic-specific effects and potential learning or fatigue effects across experiments, enhancing the robustness of the research findings.

For the four group mentioned above, the experimental procedure for the two research topics, *HotpotQA* and *CNN/Daily*, is as follows:

- **Step1 - Description (10 minutes):** Explain to participants the module's function and the meanings of nodes and edges. This explanation will also appear at the top of the displayed **UI**.
- **Step2 - Hands-on Research assistant interface (75 minutes):** Experiment and group setting mentioned above.
- **Step3 - Questionnaire (15 minutes):** To obtain a more intuitive grasp of participants' experiences with the interface, we designed a comprehensive survey questionnaire covering the entire interface, as illustrated in the table below. Participants will complete a survey questionnaire and offer constructive suggestions for improving the Research Assistant Interface.
- * To assess participants' learning outcomes, we let them create top-down survey route diagrams during experiment 2, using the example provided in **Figure 8.1** as a guide. Due to time constraints, participants only recorded knowledge they could readily comprehend in the diagram. The diagram comprises three main characteristics:

    1. A top-down conceptual structure that progresses from abstract concepts to specific details.

    2. Key-points showing paper relevance identified through the relevance-based summary view.

    3. Details learned from the *diff-table* to enhance understanding.

- * The experiment is planned to spend 100 minutes, with 10-minute breaks between steps, totaling 120 minutes.

8.2 Subjective evaluation

We gather basic information from experiment participants, including their academic year, frequency of conducting research surveys, and familiarity with the two research topics used in the experiment. During the experiment, we assess participants' survey effectiveness by having them read the views presented in the interface and create their own survey logic diagrams. Afterward, we collect participants' subjective satisfaction ratings and improvement suggestions for the entire system through questionnaires.

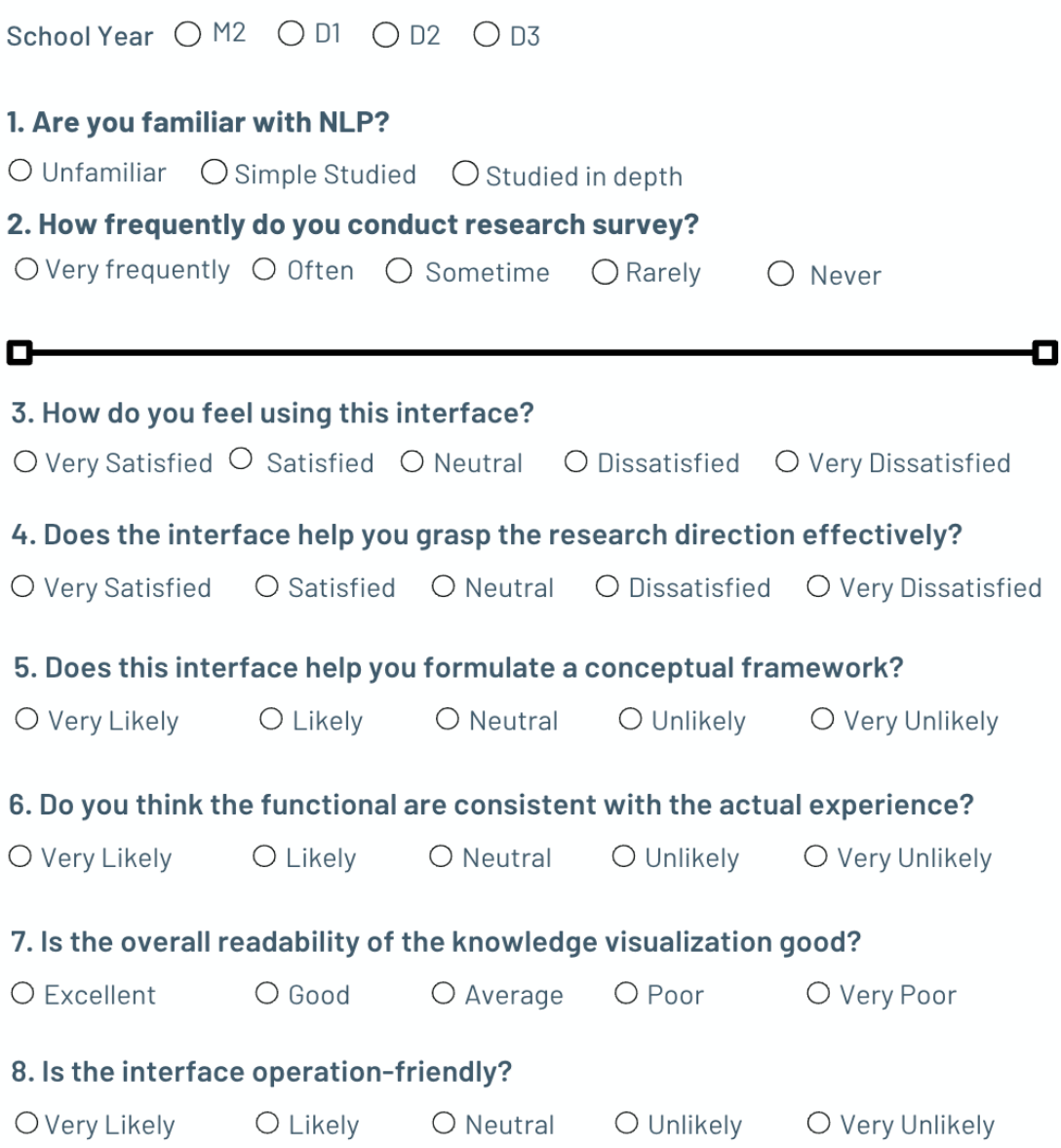


Figure 8.2: Questionnaire of subjective system evaluation

8.2.1 Overall evaluation

The questionnaire for the overall system evaluation is shown in **Figure 8.2**. The **Table 8.2** summarizes the overall evaluations of the research survey assistant interface from 11 experimental participants. We assessed user satisfaction with this interface using a 5-point scale, based on responses to questions 3-8 in the questionnaire shown in **Figure 8.2**. Scores ranged from 5 (highest satisfaction) to 1 (lowest satisfaction). The abbreviations in **Table 8.2** are detailed as follows:

- **SY:** School years.
- **G:** Group of experiment.
- **FAM:** Familiarity with the **NLP** field.
- **FRE:** Frequency of conducting research survey via reading academic papers.
- **SAT:** Overall Satisfaction of the interface.
- **RD:** Effectiveness of grasp the research direction.
- **FC:** Effectiveness of formulating a conceptual framework.
- **CF:** Consistency of functional descriptions.
- **RA:** Overall readability fo knowledge visualization.
- **OPE:** Satisfaction of **UI** Operation.

In **Sections 8.2.2 - 8.2.4**, we present detailed subjective experimental results for each subsystem. **Sections 8.2.5** discusses advantages of the entire research survey assistant interface compared to independent subsystems.

The results show that novice researchers at various stages (from master's to PhD students) generally expressed satisfaction with the research survey assistant interface according to the metrics mentioned in **Section 8.2.1**. In addition, by analyzing the variance in evaluation results across different metrics from the 11 participants and collecting their actual experiences using the system, we know that participants' familiarity with the **NLP** field, school year, and frequency of conducting research surveys did not significantly impact their experience with this system. Instead, the main factors contributing to differences in system experience are:

- **Users' preferred research survey route:** Some participants, while able to gain directional guidance from the top-down survey route, found the initial presentation of concise, broad concepts from the fish-bone diagram confusing. They questioned whether these concepts had practical solutions. These participants often preferred a bottom-up approach: first examining specific details in the diff-table to gain concrete insights, then expanding their thinking to more abstract

Table 8.2: Overall system evaluation
***Simple Studied :** Has taken NLP courses, but no experience in Question Answering and Text Summarization

No.	SY	G	FAM	FRE	SAT	RD	FC	CF	RA	OPE
1	M2	1	Simple Studied	Sometime	4	4	4	3	3	4
2	D2	1	Unfamiliar	Sometime	4	4	4	4	5	5
3	D3	1	Unfamiliar	Rarely	4	4	3	5	4	4
4	D3	2	Unfamiliar	Often	3	4	4	3	3	2
5	M2	2	Unfamiliar	Sometime	4	4	4	4	4	3
6	D2	2	Unfamiliar	Rarely	4	4	3	4	5	3
7	D1	3	Simple Studied	Often	4	4	4	4	4	3
8	D2	3	Simple Studied	Very frequently	4	3	3	4	4	4
9	D2	3	Simple Studied	Very frequently	5	5	5	5	4	5
10	D1	4	Unfamiliar	Rarely	4	4	4	4	3	3
11	D2	4	Unfamiliar	Often	4	4	3	5	4	4
Avg	-	-	-	-	4.0	4.0	3.72	**4.09**	3.91	3.64
Var	-	-	-	-	0.20	0.20	0.42	0.49	0.49	**0.85**
Std	-	-	-	-	0.20	0.20	0.65	0.7	0.7	**0.92**

concepts. For these researchers, presenting the **relevance-tree** and fish-bone diagram after initial exploration of the diff-table provided more effective research assistance. In contrast, participants accustomed to top-down logic adapted well to the system's visualization methods and inter-view connections. Their habitual use of a top-down survey approach meant that the system's knowledge overview and exploration of knowledge associations significantly enhanced their understanding of new research topics.

- **Users' preferred research domain:** The 11 participants were unfamiliar with the *HotpotQA* and *CNN/Daily* topics used in our experiments. However, their diverse specializations within computer science led to varying research habits. Some researchers, focused on theoretical modeling without application development, showed less interest in the concrete development-related elements presented in the diff-table and **relevance-tree**. Their limited background in areas like datasets and pre-training dampened their motivation to explore these aspects. In contrast, researchers engaged in application development and engineering-related research found the system's approach of concretizing broad concepts more appealing. Despite potential

unfamiliarity with **NLP**, these participants were keen to delve into the system's viewpoints and issue ontology.

we can glean the following insights from detailed metrics in **Table 8.2**.

- Most participants gained a sense of research direction **(RD)** during the experiment, enabling them to break down tasks within abstract concepts. However, regarding the establishment of a logical conceptual framework **(FC)** through brief learning , some researchers faced challenges. While they could intuitively grasp the basic elements of the research topic, the abundance of cross-connection of research elements requiring in-depth exploration from the relevance tree to the ***diff-table***—made it difficult to establish a deep conceptual framework without further refinement of the internal logic.
- The actual user experience in total system aligned with the basic functions the system aimed to convey **(CF)**. However, the readability **(RA)** evaluation metric showed larger variance in participants' assessments. This variance primarily stemmed from diverse opinions on the comprehensibility of the three views, which will be elaborated upon in **Sections 8.2.2 – 8.2.4**.
- The operational interaction experience **(OPE)** among the three UIs showed larger variance in participants' evaluations. Those familiar with bottom-up research methods expressed particular concern about the absence of **UI** operational interaction flowing from ***diff-table*** to relevance tree to ***fish-bone*** diagram. Additionally, some participants emphasized the importance of allowing users to control the amount of survey information presented based on their specific needs.

The participants' individual evaluations of each subsystem will provide deeper insights into how their survey habits and their research domain influenced the experimental results.

8.2.2 Subjective Evaluation on Fish-bone

The evaluation results for the Subsystem: **Relevance-tree** are presented in **Table 8.3**. Overall, participants provided favorable evaluations across the three assessment criteria. Factors influencing users' ratings for the fish-bone diagram correlated with their preferences for this survey approach.

we can glean the following specific insights from detailed metrics in **Table 8.3**. and participants' feedback.

Table 8.3: Subsystem evaluation : ***Fish-bone***
Functional consistency → Cons.
Survey continuity → Cont.
Comprehensibility → Comp.

No.	**SY**	**G**	**FAM**	**FRE**	**Cons.**	**Cont.**	**Comp.**
1	M2	1	Simple Studied	Sometime	4	3	3
2	D2	1	Unfamiliar	Sometime	4	4	3
3	D3	1	Unfamiliar	Rarely	4	5	5
4	D3	2	Unfamiliar	Often	3	4	4
5	M2	2	Unfamiliar	Sometime	4	4	4
6	D2	2	Unfamiliar	Rarely	2	3	3
7	D1	3	Simple Studied	Often	3	4	4
8	D2	3	Simple Studied	Very frequently	2	2	2
9	D2	3	Simple Studied	Very frequently	5	5	5
10	D1	4	Unfamiliar	Rarely	3	4	4
11	D2	4	Unfamiliar	Often	4	4	5
Avg	-	-	-	-	3.45	**3.81**	**3.81**
Var	-	-	-	-	0.79	0.69	**0.88**

- **Researcher No.9** possesses foundational knowledge in **NLP** from university courses and regularly engages with research literature. Though he hasn't directly applied the research directions outlined in the ***fish-bone*** diagram, his basic grasp of **NLP** concepts, such as Question Answering, text summarization—allows him to appreciate how the diagram's task and issue information reflects current research trends. Notably, the hierarchical overview derived from actual paper data in the ***fish-bone*** diagram effectively illuminates the main research directions for him.
- **Researcher No.3**, unfamiliar with **NLP** and infrequent in paper-based research surveys, finds the ***fish-bone*** diagram's concepts particularly stimulating when approaching new research topics. He prefers beginning with abstract, easily digestible general points before delving deeper and extrapolating specific details. This top-down thinking model aligns perfectly with his cognitive approach, effectively guiding his further exploration.
- In contrast, **Researcher No.8**, despite having **NLP** experience, didn't evaluate the ***fish-bone*** diagram positively. He favors a bottom-up survey approach, focusing on specific problems rather than broad con-

cepts. While the ***fish-bone*** diagram offers some hierarchical knowledge structure, its abstract concepts alone don't provide enough context for him to grasp the detailed research background. Consequently, it failed to significantly enhance his survey efficiency. Although **Researcher No.6** lacking of **NLP** background, he also prefers a bottom-up approach. Thus, the fish-bone view offered him little benefit.

8.2.3 Subjective Evaluation on Relevance-tree

The evaluation results for the Subsystem: **Relevance-tree** are presented in **Table 8.4**. Participants' evaluations of the **relevance-tree** were largely consistent. Most found that its visualization approach, combined with the summary content in nodes and edges, offered significant insights into the relationships among multiple articles. The summaries for 'Issue Resolved' and 'Issue Finding' were praised for their specificity and clarity. This provided them with a foundation to extent on the abstract concepts presented in the ***fish-bone*** diagram. Consequently, most participants awarded high scores for the comprehensibility criterion. However, when it came to the chain-like logic of 'Issue Finding' → 'Issue Resolved' → 'Issue Finding' as a framework for deducing potential novel directions in this research topic, participants generally struggled to assess its validity. Moreover, they found it challenging to employ divergent thinking based on the view's summaries to construct an ideal path of innovation. There are two main reasons for this issue:

- The classification errors discussed in **Section 6.4.2** led to a small number of summaries failing to accurately reflect their corresponding issue ontology. This inaccuracy indirectly compromised the validity of certain issue chains.
- Assessing the potential novelty of a research branch with limited information and time is inherently difficult. Even with the aid of ***fish-bone*** diagram and ***diff-table***, novice researchers struggle to conceptualize such complex, multidimensional-knowledge structure reflect research originality.

8.2.4 Subjective Evaluation on Diff-table

The evaluation results for the Subsystem: ***diff-table*** are presented in **Table 8.5**. The results indicate that most participants find the ***diff-table*** effective in providing specific summary information under commonly agreed view-

Table 8.4: Subsystem evaluation : Relevance tree

No.	SY	G	FAM	FRE	Cons.	Cont.	Comp.
1	M2	1	Simple Studied	Sometime	3	4	5
2	D2	1	Unfamiliar	Sometime	4	4	4
3	D3	1	Unfamiliar	Rarely	2	5	5
4	D3	2	Unfamiliar	Often	2	2	5
5	M2	2	Unfamiliar	Sometime	2	2	4
6	D2	2	Unfamiliar	Rarely	3	4	3
7	D1	3	Simple Studied	Often	2	4	4
8	D2	3	Simple Studied	Very frequently	2	4	4
9	D2	3	Simple Studied	Very frequently	4	5	4
10	D1	4	Unfamiliar	Rarely	1	4	5
11	D2	4	Unfamiliar	Often	2	4	3
Avg	-	-	-	-	2.45	3.81	**4.18**
Var	-	-	-	-	0.79	**0.88**	0.51

points. They also gain valuable insights by comparing commonalities and differences across multiple articles. These summaries, organized according to each article's structure, boost their motivation for survey continuity. However, **researchers No.4, No.7, and No.10** found the *diff-table*'s content less helpful. The specific reasons are as follows:

- **Researcher No.4**'s feedback revealed: His limited grasp of **NLP** applications hindered his ability to quickly connect with the viewpoints. Consequently, he struggled to compare similarities and differences across multiple articles using viewpoints that delve into specific details like datasets, pre-training, and baselines.
- **Researcher No.7 and No.10's** feedback revealed: In experiment 2, some viewpoint summaries for articles selected from the **relevance-tree** were not fully displayed, particularly the limitation and future work sections. This occurred due to two main reasons: (1) the original article did not mention limitation or future work-related content, or (2) sentences reflecting the viewpoint did not contain the keywords described in **Section** 7.3.2, leading to a failure in extracting the original sentences.

Table 8.5: Subsystem evaluation : *Diff-table*

No.	SY	G	FAM	FRE	Cons.	Cont.	Comp.
1	M2	1	Simple Studied	Sometime	4	4	5
2	D2	1	Unfamiliar	Sometime	3	4	5
3	D3	1	Unfamiliar	Rarely	4	4	4
4	D3	2	Unfamiliar	Often	2	4	2
5	M2	2	Unfamiliar	Sometime	3	4	4
6	D2	2	Unfamiliar	Rarely	5	5	5
7	D1	3	Simple Studied	Often	2	3	2
8	D2	3	Simple Studied	Very frequently	4	4	4
9	D2	3	Simple Studied	Very frequently	4	5	3
10	D1	4	Unfamiliar	Rarely	2	4	3
11	D2	4	Unfamiliar	Often	4	4	4
Avg	-	-	-	-	3.36	**4.09**	3.81
Var	-	-	-	-	0.96	0.26	**1.06**

8.2.5 Top-down process *VS* independent subsystem

Table 8.6 compares the evaluation results of the three independent subsystems (Baseline: experiment 1), where participants were unaware of the triggers between views, with the improved understanding of the entire research route when triggers are added (experiment 2). Participants generally found that experiment 2, which incorporated triggers, better reflected the functional consistency of each component in the top-down survey route. This improvement stems from the presence of relevant connections between views. In contrast to the isolated view experience in Experiment 1, these connections enable participants to grasp the origin of each view, thereby providing more explicit background information. For instance, the issues and tasks in the ***fish-bone*** diagram can be expanded into summaries of multiple related papers. The papers in the **relevance-tree** all fall within a specific task in the ***fish-bone***, while the issues in the ***fish-bone*** further subdivide this task into various research branches. Consequently, this top-down linking logic aids novice researchers in comprehending each independent view more effectively.

Regarding Survey continuity enhancement, **researchers No.4, 7, and 8** note that while the system provides a connection basis among multiple views, it lacks detailed evidence to refine inter-view links. For instance, the related issue connecting the ***fish-bone*** and **relevance-tree** is merely a brief

Table 8.6: System evaluation : Top-down survey process (experiment 2) - Compare with independent subsystem

No.	SY	G	FAM	FRE	Cons.	Cont.	Comp.
1	M2	1	Simple Studied	Sometime	+ 2	+ 2	+ 2
2	D2	1	Unfamiliar	Sometime	+ 2	+ 2	+ 1
3	D3	1	Unfamiliar	Rarely	+ 3	+ 3	+ 2
4	D3	2	Unfamiliar	Often	+ 2	0	+ 1
5	M2	2	Unfamiliar	Sometime	+ 1	+ 1	+ 1
6	D2	2	Unfamiliar	Rarely	+ 2	+ 1	+ 1
7	D1	3	Simple Studied	Often	+ 1	0	+ 2
8	D2	3	Simple Studied	Very frequently	+ 1	0	+ 2
9	D2	3	Simple Studied	Very frequently	+ 2	+ 2	+ 3
10	D1	4	Unfamiliar	Rarely	+ 1	+ 2	+ 1
11	D2	4	Unfamiliar	Often	+ 2	+ 2	+ 2
Avg	-	-	-	-	**+ 1.72**	+ 1.36	+ 1.64
Var	-	-	-	-	0.38	**0.96**	0.41

summary, without point-to-point references to specific paragraphs in papers. This absence of original text references hinders their ability to judge the reasonability of subsequently generated content.

Participants gave high scores for improved view readability, regardless of their primary research fields or survey habits. They generally believed that combining abstract concepts with summaries of specific content in a hierarchical structure enhanced their comprehension. The experiment 2 improved their ability to logically connect knowledge and construct corresponding research conceptual frameworks.

8.2.6 Outcome and Bias analysis

Following the example in **Figure** 8.1, participants created top-down survey route diagrams using the top-down process view (experiment 2) in the research survey assistant interface. This section first evaluates their learning outcomes based on the following criteria:

- **Sufficient:** 3-point scale, ranging from content-rich to content-poor **[+3 — +1]**.
- **Reasonableness:** 3-point scale, ranging from logical to illogical outcomes **[+3 — +1]**

Table 8.7: Evaluation of learning outcomes
***APT** : Adaptability to top-down processes (feedback from participants)

No.	SY	G	FAM	FRE	APT	Sufficient	Reasonableness
1	M2	1	Simple Studied	Sometime	Good	3	2
2	D2	1	Unfamiliar	Sometime	Normal	2	2
3	D3	1	Unfamiliar	Rarely	Good	2	3
4	D3	2	Unfamiliar	Often	Normal	1	1
5	M2	2	Unfamiliar	Sometime	Normal	3	1
6	D2	2	Unfamiliar	Rarely	Bad	1	2
7	D1	3	Simple Studied	Often	Good	2	2
8	D2	3	Simple Studied	Very frequently	Bad	2	2
9	D2	3	Simple Studied	Very frequently	Good	3	2
10	D1	4	Unfamiliar	Rarely	Normal	2	1
11	D2	4	Unfamiliar	Often	Good	2	2

Participants demonstrated different learning outcomes based on various characteristics, including their familiarity with the **NLP** field, Adaptability to top-down processes, and knowledge of data science engineering. **Table 8.7** presents the evaluation of learning outcomes of 11 participants, while **Table 8.8** uses correlation analysis to illustrate how different participant characteristics influenced these outcomes. The code of correlation analysis is shown in **Code B**.

From the **Table 8.7**, we realize main factors affecting their learning outcomes are shown below:

- **Adaptability to top-down processes(feedback from participants):** A strong positive correlation emerged between participants' adaptability to the top-down process and their learning outcomes, demonstrating how learning style significantly affected their diagram quality. Researchers who scored 'BAD' in **APT** struggled to expand abstract concepts into specific details and verify underlying information. Their top-down survey route diagrams reflected these difficulties through fewer branches and noticeable hesitation when confronting ambiguous logical structures under time constraints.
- **Familiarity with the NLP field:** Familiarity with the **NLP** field also demonstrated a strong positive correlation with participants' learning outcomes. Participants who were unfamiliar with **NLP** struggled

Table 8.8: Result of correlation analysis - Spearman method [130]

	Sufficient	**Reasonableness**
G*(Group)*	-0.057	**-0.313**
FAM	**0.462**	0.280
FRE	0.199	-0.010
APT	**0.480**	**0.397**

with their survey efficiency due to limited knowledge of the field's research methodologies. While these participants could grasp and expand upon abstract concepts to some degree, they had difficulty with specific technical terminology. This led to incomplete top-down survey route diagrams that lacked detailed exploration.

- **Frequency of conducting research survey via reading academic papers:** Participants who frequently conduct research surveys produced more sufficient top-down survey route diagrams. This may be attributed to their habit of reading numerous articles and recording key research points, which motivated them to be more proactive in creating top-down survey route diagrams.
- **Experiment sequence:** The order of experiments influenced learning outcomes. Groups 1 and 2, who completed Experiment 1 first, gained familiarity with the system and produced more logically coherent top-down survey route diagrams in Experiment 2. In contrast, Groups 3 and 4, who began with Experiment 2, had limited time to adapt to the top-down process. This resulted in diagrams that showed weaker coherence, especially in their understanding of connections between subsystems.

Due to the limited sample size in this experiment, the analysis results may contain biases in the following aspects. Future large-scale supplementary experiments will be needed to enhance the stability of the statistical analysis.

- **Data Distribution Bias:** The data distribution in **Table 8.7** reveals several imbalances that could affect the results. For example, in the **APT** category, Good and Normal cases outnumber Bad cases. Additionally, **D2** participants make up a disproportionate share of the sample, and the **FAM** category shows a high number of Unfamiliar cases. For future experimental designs, we should incorporate preliminary testing to better balance participant characteristics and reduce potential bias.
- **Joint effect of multiple variables cause bias:** The evaluation of participants' learning outcomes may rely heavily on specific parame-

ters, potentially introducing bias into the outcome. We accumulated variables combinations to identify those that had significant effects on the learning outcome in our experiment result.
The following parameter combinations showed potential bias in the experimental results:

- **FAM:** Simple Studied, Unfamiliar
- **FRE:** Often, Rarely, Sometime
- **APT:** Bad

Researchers in this experiment who shared multiple characteristics described above demonstrated similar patterns of bias in their top-down survey diagrams.

Chapter 9

Conclusion

To effectively summarize a large number of articles within a research topic from general concepts to specific details and provide novice researchers with an appropriate top-down research route, this dissertation proposed a novel approach to research survey assistance designed specifically for novice researchers. Drawing inspiration from the top-down method of research experts, we developed an interface that guides users from broad, abstract concepts to detailed, specific information. This approach is embodied in a three-tiered visualization system—the ***Fish-bone*** diagram, Relevance tree, and Diff-table — which reduces the time and effort required for novice researchers to gain a comprehensive understanding of their field. By enabling quick familiarization with research direction from hundreds of papers and facilitating more targeted reading and questioning, our tool has the potential to accelerate the early stages of research surveys. The main contributions of this study include the following points:

1. To construct the infrastructure data for the research survey assistant interface, we develop a Compartment & Text Blocks Refinement Framework (**CTBR**). This framework identifies the internal structure of research papers in *PDF* format, accurately extracting the body text and associated sections to form the main part of the infrastructure data.
2. We mimic expert researchers' survey methods, utilizing common issue ontologies and viewpoints in academic research as foundational elements to define the requirements for our top-down survey assistance system.
3. From the infrastructure data, we extract and refine a summary dataset that exemplifies the feature of top-down survey, and we employ machine learning and prompt-engineering techniques to generate layered summaries based on this summary dataset, implement a structured top-down summary style from broad abstract concepts to specific details, based on the issue ontology and viewpoints.
4. We design and develop a comprehensive set of top-down survey sum-

marization visualization tools. These tools are arranged hierarchically from top to bottom as ***Fish-bone*** diagram, Relevance tree-structured ***KG***, and ***Diff-table*** that provide novice researchers with a more concise, clear, layered, and logical summary generated in stage 3. This approach enables researchers to quickly familiarize themselves with the outlines of hundreds of papers in a short time.

9.1 Findings

This chapter explores expansion ideas stemming from our research. The **Table 9.1** outlines the general directions.

9.1.1 Extension based on Chapter 6 : Longitudinal Insight path generation

In this dissertation, we proposed longitudinal insight survey for a research topic, establishing a tree-structured academic insight knowledge Graph based on research issues. However, this structure is complex in expressing multiple branches, still requiring researchers to spend time further exploring to understand the longitudinal overview of the entire research topic. Thus, we plan to focus on establishing a coherent insight path through the collaboration of multiple agents, providing researchers with a series of continuous research ideas and directional guidance. It has the following characteristics:

1. Issue ontology based: It is composed of issue ontology that link multiple papers, deeply mining the issue threads interspersed across multiple papers to form the environment for generating the insight path.

2. Generation mode via multi-agents: Insight path generated through the division of labor and cooperation of multiple agents, through information acquisition from the environment, analysis of the current state, dynamic programming ***(DP)***, and decision-making, to construct the comprehensible insight path.

3. Concise Representation: It presents a longitudinal survey in both temporal and spatial aspects, and provides a simpler summary representation reflecting the connections between multiple articles.

9.1.1.1 Issue Relevance

Issue relevance forms the core of longitudinal insight path generation, specifically focusing on the similarities and differences among issues in academic papers. This research discusses issue relevance from two perspectives:

Table 9.1: Finding

RQ	**Our Solution**	**Finding & extension**
Main: How to effectively summarize large amount of articles within a research topic, to provide novice researchers with appropriate top-down research route?	Develop a 'top-down' research survey assistant interface for novice researchers	→ In **Section 9.2 - 9.3** 1. Cross-interaction between subsystems 2. Enhance diversity in survey route 3. Further interaction of UI components
Sub1: How to provide novice researchers with a hierarchical bird-eyes view that reflects the causal-effect overview in a research topic?	***Fish-bone*** diagram function on reflecting logical guide chain of topic → task → issue ontology for novice researchers to navigate research topics	→ In **Section 9.3.4** Expand on the tasks and issues from the ***Fish-bone*** diagram to create a mind map resembling Research flow navigation.
Sub2: How to help researchers gain effective insights by identifying connections between the evolution and relevance across multiple research papers?	Tree-structured knowledge graph function on expanding the citation inheritance and relevance association	→ In **Section 9.1.1** 1. Retrieving longitudinal insight paths from the tree → In **Section 9.1.2** 2. A Divergent Insight View - A Survey Forest Diagram
Sub3: How to aid novice researchers in identifying similarities and differences in the research task through cross-comparison?	Abstractive summarization in a viewpoints-embedded ***diff-table*** format.	→ **Section 9.2 - 9.3** 1. Extensibility for external resource 2. Specific explanation feature for proprietary concepts

commonality (common knowledge elements reflected across multiple papers) and difference (The unique differences of each paper that emerge from multiple papers with commonality). In our research process, we expand and extend based on the work of our predecessors. The learning process involves acquiring issues known to exist in the world at that time. The learning process discussed in research papers is mostly reflected in the introduction and related work sections. These two parts are generally summaries of commonality in the learning process of that domain. However, the authors also write about the unique aspects of their papers in these two 'chapters' to reflect the novelty of their work, thus distinguishing it from other research and demonstrating their differentiation and value. Therefore, grasping this layer of relevance relationships will allow readers to better immerse themselves in the author's perspective when reading the paper, thereby quickly understanding the surface information the paper conveys. We define the Longitudinal insight path as a framework that reflects the issue relevance of multiple papers. This approach allows researchers to quickly grasp both the surface information conveyed by the papers and the implicit research directions. When presented to novice researchers, this path is expected to streamline their complex survey process during the initial stages of research, simplifying the often overwhelming task of literature review. **Figure 9.1** illustrates an example of the ideal insight path. This figure illustrates the Multi-hop task within the *HotpotQA* topic. `P1` shows the datasets involved in the Multi-hop task. `P2` expands from Multi-hop question to the Modular network approach. `P3` incorporates single-hop supporting evidence to enhance the interpretability of multi-hop question answering framework. This path showcases the commonality, differences, and inheritance patterns among similar research tasks, offering researchers a comprehensive, multi-faceted insight.

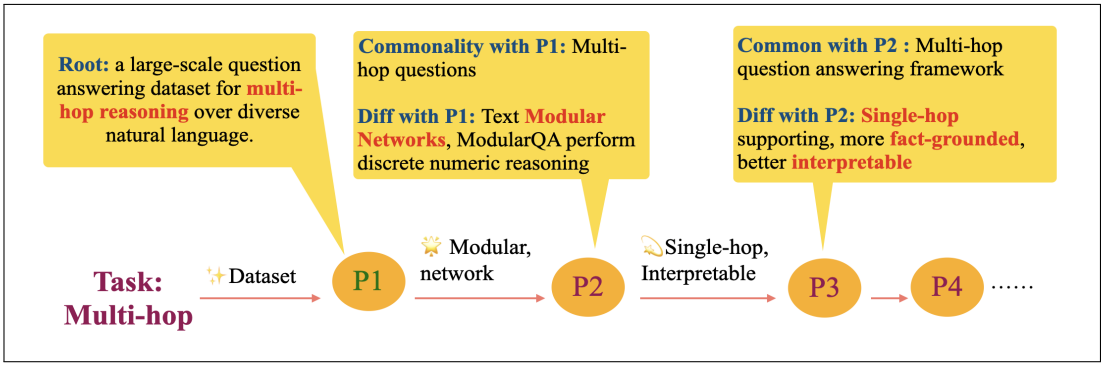


Figure 9.1: Sample insight path

9.1.1.2 Role of retrieval agent

The collaboration among multiple agents and the exchange of transparent information can enhance the precision and efficiency of the decision-making process [131]. This study defines three types of agent.

Information collector: In any learning or work process, gathering information is the crucial first step before progressing to subsequent stages. For this research, we first collect the essential information the Insight Path generation by establishing an Agent that functions as an *Information collector*. This Agent's role is to gather and convey data from the ever-evolving landscape of academic issue environment. It captures the characteristics and types of issue ontology, citation relationships between papers, and keywords of paper. The Agent then conveys this information in real-time to the *Status Analyst*.

Status Analyst: In companies, there are data analysts who parse and search for the inherent meaning of data collected by *Information collector*, and then propose constructive strategies to report to decision-making superiors. Drawing a parallel to this research, the *Status Analyst* performs classification learning on the implicit category information carried in the issue ontology, determining whether there are strongly related issue ontology and their corresponding papers in the environment. If found, these are then handed over to the *Decision maker* for adjudication.

Decision maker: The *Decision Maker* receives proposals from the analyst and makes judgments based on previous experience. The outcome may be full acceptance, partial acceptance, or rejection of the proposal. In this study, we establish a *Decision Maker* agent that makes decisions based on its goal of orientation. This orientation aims to maximize total benefits within budget constraints. As current benefits and external factors may alter previous strategy of decisions, the *Decision Maker* optimizes the entire system to determine how to select issue ontology to construct insight path and reflect their differences.

9.1.1.3 Insight path generation

This section explores how multiple agents collaborate to collect, analyze, and identify key elements forming the insight path from complex environmental variables. We also illustrate how these agents dynamically generate and determine the path's expansion direction. **Figure 9.2** illustrates the Multi-agents' workflow of insight path generation.

A. Environment: Human-environment involve complex interdependencies of agents that need to be taken into account when modeling these

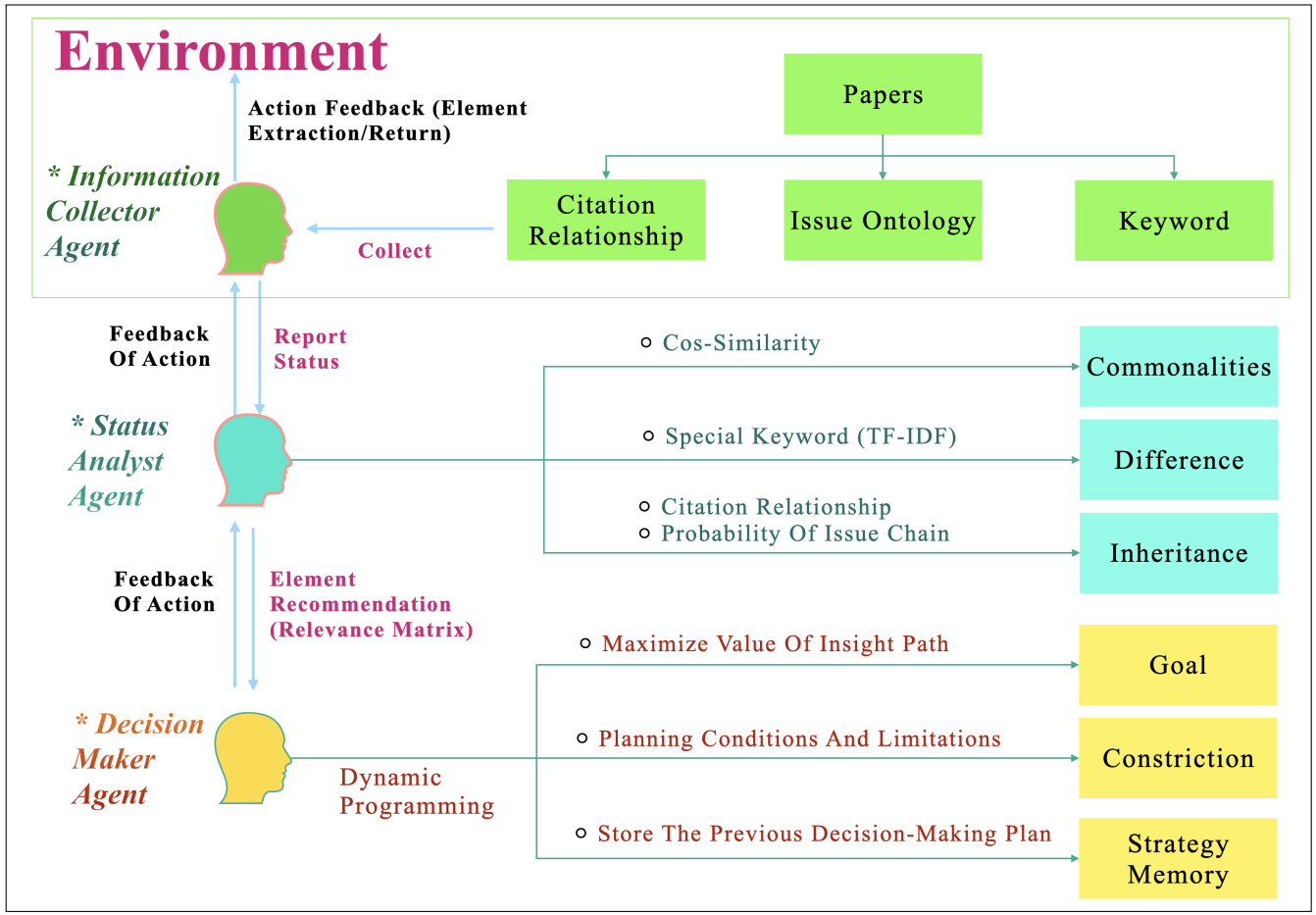


Figure 9.2: Multi-agents' workflow of insight path generation.

interaction [132]. The environment we establish integrates both the complex internal characteristics of academic papers include issue ontology sentence, citation relationship, and keywords in paper, as illustrated in **Figure 9.2**. The *Information collector* needs to gather and explore key elements for generating the insight path, which include:

1) Issue ontology: Machine learning predicts which category of issue ontology these insight sentences belong to, outputting probability values (fitness) to the *Information collector* agent. This research examines 'Resolved issue' and 'Finding issue' as an example. Within a specific topic's paper environment, the *Information collector* agent gathers the probability values (fitness) to these two types of issue ontology.

2) Citation relationship: Citation relationships exist among papers in the research environment, supporting the inheritance structure of multiple studies. Authors cite articles with specific purposes: to use tools or ideas from previous papers, reference ideas, or extend prior work. Consequently, papers with citation relationships often reveal insights into inheritance in research approaches.

3) Keyword: When comparing papers, each one inevitably presents unique content to highlight its differences. The keywords reflecting this unique information become an essential part of the research environment. For each selected paper, keywords are calculated using *TF-IDF*. The set of selected papers is dynamic, changing in real-time based on the Decision maker agent's ongoing decisions. These keywords form a crucial component of the insight path's representation.

*: Agents must coordinate and collaborate to navigate this complex environment. The specific roles and responsibilities of each agent are as follows:

B. Information collector - Environmental Intelligence Agent: The *Information collector* gathers elements and dynamics from the current environmental stage. It first obtains information on candidate issue ontology and corresponding papers (noting that some elements will be added to the insight path and removed from the environment). It then collects sentence-data on issue ontology probabilities, paper keywords, and citation relationships. This collected information is passed to the *Status Analyst* for further analysis and element filtering.

C. Status Analyst - Relevance matrix: The *Status Analyst* agent's primary role is to conduct insightful analysis on commonality, differences, and inheritance. This analysis aims to uncover deep connections hidden beneath the surface data of multiple articles.

1)Commonality: To maintain a common conceptual thread among the issue ontology elements in the insight path, Status Analyst agent computes the cosine similarity between candidate Issue ontology in our environment and the overall issue ontology in the existing insight path, as illustrated in **Equation 9.1**. Here, $v_{commonality}$ represents the value (degree) of the commonality, **ca** represents the candidate issue ontology, while *path* denotes the issue ontology currently present in the insight path.

$$\mathrm{v}_{\mathrm{commonality}}(\mathbf{ca}, path) = \cos(\mathbf{ca}, path) = \frac{\mathbf{ca} \cdot path}{|\mathbf{ca}| \mid path \mid} \tag{9.1}$$

2)Difference: Additionally, the Status Analyst tries to identify differences among papers. Instead of sentence similarity, it aims to highlight the uniqueness of candidate issue ontology compared to those in the current insight path, effectively showcasing differences of unique words. It calculates cosine similarity based on word distances rather than sentences due to some sentences with same issue ontology often share similar grammatical structures, whereas using keywords as units more effectively highlights distinctive features. **Equation 9.2** expresses the method for calculating the differences of candidate issue ontology.

$$\mathrm{v}_{\mathrm{diff}}(\mathbf{ca}, path) = 1 - \cos(\mathbf{keyword_{ca}}, keyword_{path}) \tag{9.2}$$

3)Inheritance: To reflect inheritance, first determine whether the candidate issue ontology and the tail-end issue ontology in the insight path form an issue chain (Issue Finding $\rightarrow$ Issue Resolved). The probability of this issue chain is determined through machine learning, based on the probabilities of Issue Finding and Issue Resolved. Moreover, if the papers in the issue chain have a citation relationship, it's inferred that the reliability of inheritance is enhanced, and appropriate weighting is applied in the calculation. The calculation method is illustrated in **Equation 9.3**, with the results incorporated into the Inheritance component of the relevance matrix. Here, v_{chain} represents the constituent value of the issue chain, $proba_{Finding}$ and $proba_{Resolved}$ denote the probability values of these two issue ontology classified through machine learning, and gamma indicates the degree of citation influence.

$$V_{chain} = \gamma * (proba_{Finding} * proba_{Resolved}) \tag{9.3}$$

The results from **C.1-C.3** are then incorporated into the relevance matrix as elements, reflecting their relationship with the current state of the insight path. This Relevance matrix updates in real-time to reflect changes in the dynamic environment.

D. Decision maker: Dynamic programming *(DP)* The *Decision Maker* agent plans and decides based on the relevance matrix provided by the Status Analyst, it try to balance immediate judgments with long-term goals. The immediate goal is to make the optimal decision in the current environment by selecting new issue sentences. These sentences either become new nodes in the path or replace existing ones, maximizing short-term benefits. It attempts to maximize the value of the insight path while controlling the cost required to generate this value. Thus, we implement this process using the knapsack method of ***DP*** shown in **Equation 9.4,9.5 and 9.6**. Here, x_i expresses the sequence of issue ontology in the insight path, α and β expresses the acceptable range of differences within the insight path. The constraints condition of ***DP*** represent the *Decision Maker* agent's need to control decision reliability. Specifically, the agent aims to extract sentences from a set of issue ontology with strong commonality, while choosing those that differ significantly from the current issue ontology. This constraint encourages the *Decision Maker* agent to avoid selecting sentences that overlap excessively with previous content, thus enhancing the novelty and directional guidance of the generated path.

$$Max. \qquad f(x) = \sum_{i=0}^{n} (v_{commonality} + v_{chain}) x_i \tag{9.4}$$

$$Subject \quad to. \qquad \frac{\sum_{i=1}^{n} v_{diff} * x_i}{n} \in (\alpha, \beta) \tag{9.5}$$

$$Subject \quad to. \qquad \sum_{i=1}^{n} x_i <= Max.length_{insight\_path} \tag{9.6}$$

The *Decision Maker* agent stores the strategy and value from each instantaneous decision in its memory. After several iterations, the *Decision Maker* agent compares the accumulated value from each epoch and chooses the path with the highest overall benefit as its final decision.

Using issue sentences as the primary elements of the insight path may include redundant expression that cause researchers to lose interest. Thus, in the insight path visualization stage, we aim to retain crucial information of issue ontology while concisely summarizing the text. We achieve this

through prompt engineering using *gpt-4o-mini*. The sample of few-shot prompt description is as follows:

```
prompt = f""" Your task is to make a brief summary that
    reflects the point of issue this work solved.

     <Original text>: For future work, we are interested in
    exploring scaling up our method and other scenarios, e.g.,
     commonsense reasoning (Talmor et al., 2022) and
    biomedical retrieval (Nentidis et al., 2020; Zhang et al.,
     2022b)..

     <Summary>: Exploring  commonsense reasoning and
    biomedical retrieval scenarios.

     Please summarize the following text:

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

Listing 9.1: Sample Prompt - Summary of Issue sentence

We plan to invite research expert in a specific research topic to create gold standard for the imagined insight path. The gold standard setting follow the criteria description bellow:

(1) Based on the previous experience to simulate the route that reflect how to achieve the survey process.

(2) Concentrate on the facts that not related to the future direction of that research field, considering their specific characteristics of common and diff, and ignore the part of analysis and detail explanation.

We plan to objectively evaluate the effectiveness of the insight path from the following 4 aspects:

1.Consistency: The factual consistency between the summary and the original source.

2.Correctness of issue ontology: Whether the summary content containing corresponding issue ontology correct.

3.Comprehensible: The expression of summarization in the nodes, whether the reader can understand the general meaning of the content and find the key points of the survey that directly.

4.Sufficient Coverage: Whether the important information that directly reflects the viewpoints of the sentence has been fully expressed.

To validate the effectiveness of the insight path in real-world scenarios, we must account for subjective biases among researchers. We plan to invite researchers of various levels to evaluate whether the generated insight path enhances the quality of longitudinal surveys. Our planned evaluation metrics

are as follows:

1.Comprehensible of abstractive summarization: Whether researchers can comprehend and contrast the commonality and differences presented in the abstractive summarization of multiple papers.

2.Logical coherence of relevance descriptions: The presentation style of the Insight path, assessing its effectiveness in enabling researchers to continually perceive logical connections between issues across multiple articles

3.Guidance-able: whether researchers can independently analyze the inheritance, commonality, and differences among multiple papers via insight path.

we make a proposal for make an extension on **Chapter 6**. It calls researchers' longitudinal insight surveys by establishing research insight paths using dynamic programming through multi-agents based on issue ontology. We first use machine learning methods to categorize issue ontology, then set up the environment and objective function based on the probability distribution and commonality of issue ontology. To reflect the differences between elements in the insight path, we incorporate differential feedback in weight settings and constraints. Finally, we use multiple agents to implement dynamic programming and generate optimized insight paths.

9.1.2 A Divergent Insight View - A Survey Forest Diagram

While the longitudinal insight view demonstrates citation inheritance and development across multiple papers, the overall direction of these citations remains unclear. A more effective approach would be to provide explicit indicators of citation clues. For example, we could create a diagram where multiple well-supported branches diverge from a single knowledge source, based on original citations. This would allow novice researchers to use a central point as a foundation, encouraging divergent thinking and enabling them to construct a branching logical framework of research topic more effectively. Therefore, this extension aims to assist novice researchers in understanding the divergent insights from multiple papers on the same research topic. Starting with a survey paper as the survey root, it creates a divergent-thinking forest diagram based on three characteristics of citations: intentions, motivations, and clues, to help expand the survey perspective for novice researchers.

We try to discover divergent directions of survey papers by defining the characteristics of their citations. We create multiple layers and embed the

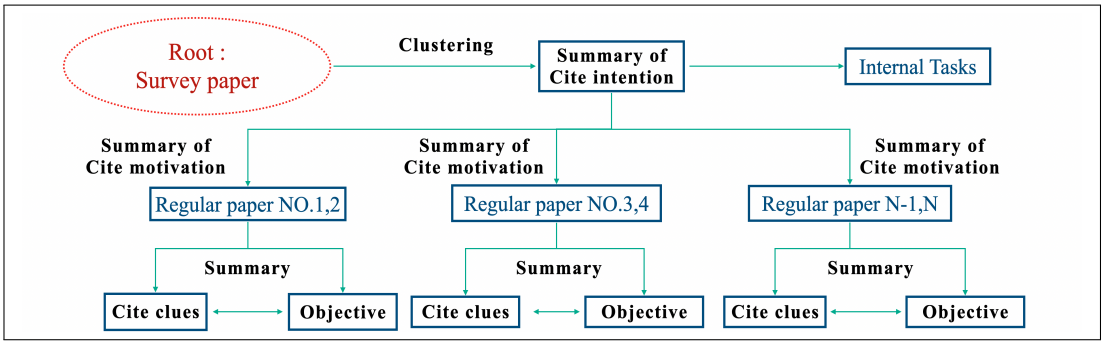


Figure 9.3: Survey forest : Strategy Overview

corresponding summaries that reflect these citation characteristics in each layer. Finally, we construct a survey forest diagram by linking survey papers, cited regular papers, and their summaries of the citation characteristics. The definition of Divergent Insight Survey and its internal concepts are as follows:

(1) Divergent Insight Survey: Divergent thinking is a cognitive style that facilitates idea generation in situations with vague selection criteria and multiple correct solutions, emphasizing mental flexibility [133]. In academic surveys, insights from divergent guidance can offer novice researchers helpful hints. This enables them to conduct more in-depth surveys from various perspectives and directions.

(2) Survey Forest Diagram: The forest diagram displayed results from paired observations and events for the similar article of feature, along with overall effects [134]. There are multiple survey papers on a research topic, and each survey paper cites multiple regular papers. They may have related research purposes and directions, and the characteristics of the forest diagram will adapt to the expression of such multi-directional, multi-branch overall effects.

(3) Survey Paper & Regular Paper: A survey paper is organized by experts in the field to provide background knowledge, related tasks, and future direction speculations. It systematically presents an overview of the specific research topic. A survey paper can be considered a root. Combining this root with the papers cited in various sections can connect their citation logic to form a citation clue.

(4) Citation intention: The authors of a survey paper conduct extensive literature research on a particular topic and organize the documents

into sections based on research task segmentation. Hence, the citations in each section embed the author's intention, reflecting the author's direction of citation.

(5) Citation motivation: The authors of survey papers usually indicate their motivation in the in-text citations. Similar to citation sentiment, this explains why the author cites a particular paper [135]. The text embedded in and near the citation often reflects the author's citation motivation.

(6) Citation clues: To determine the connection between a cited work and the citing document, analyze clues from the latter's author. Deep-mining on the citation's purpose, function, and motive is essential for measuring the work's impact [136]. This study's citation clue consists of citation intention and motivation. Creating a summary that reflects the citation clue can show the logical structure between the survey paper and the papers it cites more intuitively. This assists novice researchers in diverging on key research points logically.

Based on the above concepts, we integrate a strategy to support the Divergent Insight Survey in multiple layers. We try to demonstrate the Divergent Insight View through the following process:

- **1.** First, using the research topic of *HotpotQA* as an example, we extract survey papers with the keyword – *HotpotQA* in the infrastructure data, along with the regular papers they cite, to form the basic prototype (nodes and edges) of the forest diagram.

- **2.** Next, we locate the citation content based on the section where a regular paper is cited in a survey paper and pinpoint the location of the citation (which sentence is cited in the text). To make the diagram structure more concise, we cluster the citation content to present multiple tasks based on the citation content and expand into the citation intent.

- **3.** Then, we use prompt engineering combined with citation content to create abstractive summarization highlighting regular papers' citation clues. This showcases the citation clues of each regular paper within the entire forest, forming the extension nodes of each paper.

- **4.** Finally, we extend another edge from the regular paper within each

diagram to output its research objective, and provide a summary of the association between the cite clue and the research objective, specifically expressing the connection between survey papers and regular papers. The strategy overview is shown in **Figure 9.3**. Throughout the diagram, we use prompt engineering with **LLM** – *GPT-4o-mini* to achieve abstractive summarization for each part. The sample few-shot prompt description for 'Summary of Cite Motivation' is shown as follows:

```
prompt = f""" Your task is to make a brief summary that
    reflects the reason the author do the in-text citation.

     <Original text>:Although automatic data collection can
    obtain large-scale examples,it is restricted to limit
    reasoning types dependent on the designed heuristic
    methods [11].

     <Summary>: Automatic data collection limits reasoning
    types based on heuristic methods.

     Please summarize the following text:

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

Listing 9.2: Sample Prompt: Summary of Cite Motivation

This study specifically defines the direction of the Divergent Insight Survey and conducts initial framework development. This study aims to develop an in-depth Survey Forest Diagram that guides novice researchers in divergent thinking about the research topic by indicating the citation intentions among multiple papers, enabling them to quickly gain insights into potential research elements.

9.2 Limitation

9.2.1 Lack of diversity in survey route

This research presents a top-down survey approach, progressing from abstract concepts to overview summaries of relevant papers, and finally to in-depth paper analyses. However, experimental results in **Chapter 8** reveal that some Computer Science researchers prefer a bottom-up learning and exploration method. Instead of moving from abstract to concrete concepts, they tend to

start with specific content and then expand to discover the logical framework behind these details. For these researchers, beginning with abstract concepts may lead to confusion due to vague and non-specific guidance. Consequently, providing an inverse bottom-up survey route would be more beneficial for such individuals.

9.2.2 Lack of a specific explanation feature for proprietary concepts

The Research Survey Assistant interface lacks comprehensive interaction for specialized terms. Many researchers prefer to clarify specific definitions before grasping the overall design, as these precise definitions provide a solid foundation for further investigation. For such researchers, offering only sentence-level, top-down summaries fails to meet their need for conceptual expansion from specialized terms. They often seek to understand how a specific term's usage in one research topic differs from its application in other fields.

9.2.3 Lack of extensibility for external resources

This research provides text-based summaries without links to external sources. Development-oriented researchers who encounter intriguing proposals or datasets in the ***diff-table*** may wish to access related resources beyond the paper, such as dataset files, project source code, or trained models. These external resources could help researchers better assess the feasibility of the research and boost their motivation to conduct a more thorough survey.

9.2.4 Limitation in guiding research novelty

This research offers insights into novelties across multiple papers. However, novice researchers still struggle to quickly identify innovative areas while developing their conceptual understanding. They would benefit from a research navigator tool that searches for both resolved and unresolved issues within the research topic based on user-input keywords. This tool would enable them to expand their exploration of key elements using their self-developed conceptual framework.

9.2.5 limitation in applicability to other domains

The viewpoints of this research survey assistant interface was specifically developed for the **NLP** domain. When extending the system to other

computer science fields, some viewpoints—such as previous issues, objectives, limitations, and future work—remain adaptable. However, elements like pre-training, baselines, datasets, and performance metrics are **NLP** - specific. For deeper paper analysis, viewpoints must be redefined according to domain experts' consensus to match each field's characteristics. In non-computer fields like biomedicine and chemistry, domain experts must establish more appropriate viewpoints based on both their field's broad scope and specific internal features.

9.3 Future works

In **Chapter 5 - 7**, we outlined future optimization plans for each subsystem, and in **Section 9.1.1 - 9.1.2**, we presented two specific expansion directions. Building on these, this section proposes a comprehensive optimization plan for the research survey assistant interface, informed by the experimental results from **Chapter 8**.

9.3.1 Incorporate concise explanations of technical terms

Throughout the top-down survey process, our visual summaries, while concise, pose challenges for students outside the computer science field. The abundance of abbreviations for specialized terms, such as name of dataset, often requires extensive searches to comprehend. Moreover, certain technical terms carry different meanings across various disciplines. Without proper explanation, these acronyms and specialized terms can hinder the view's readability and dampen novice researchers' enthusiasm for further exploration. To address this, we plan to incorporate additional explanations for complex terms and acronyms in future iterations, thereby enhancing the summaries' accessibility and encouraging continued research interest.

9.3.2 Cross-interaction between subsystems

This dissertation provided a top-down mode trigger to link various subsystems. However, this unidirectional linking model does not adequately reflect the flexibility of surveys. Some novice researchers in the Computer Science field are more accustomed to starting by reading summaries of multiple papers to familiarize themselves with some research details, and then gradually exploring the broader directions within the field, which is a bottom-up approach to exploration. For these novice researchers, we need to design another set of bottom-up trigger modes to adapt to their needs.

9.3.3 Further interaction of UI components

While the **UI** operation style in this study is simple and user-friendly, it lacks interactivity. This limitation is evident in the potential correlations between unconnected nodes across different views. For instance, in the ***fish-bone*** diagram, similar issue ontology exist across various tasks. A desirable feature would be a function that, when clicking on one issue ontology, filters out those with similar ontology, thus enabling a partial correlation analysis for broader directions within the ***fish-bone*** diagram. Similarly, in the relevance tree, nodes in different trees may share similarities. The challenge lies in how to incorporate an additional layer of depth to this correlation analysis.

9.3.4 Research flow navigation

This research offers a top-down research perspective, enabling novice researchers to explore research routes and integrate key information to some degree. However, leveraging this research route for deeper guidance remains a future research challenge. The research mind map in the *ResearchFlow*[1] suggests the potential for efficient knowledge expansion through a workflow format. This approach would allow users to organize research elements in a multi-layered, cross-referential manner while maintaining an overview of the research topic. Building on our top-down survey conceptual model, the next major development direction will mainly involve automatically generating adjustable deep layered top-down or bottom-up survey navigation based on novice researchers' input after they've acquired initial research routes and elements.

9.3.5 Extending System Functionality Across Research Activities

This research currently focuses exclusively on the literature survey phase of academic work. Through interacting with our system, users gain a comprehensive overview of the research topic and explore logical top-down research directions. As they progress in their research activity, they can access specific views tailored to each stage, improving the efficiency of their research activities. For example, when seeking to understand the general research background, researchers can utilize the ***fish-bone*** diagram; when developing research questions and objectives, they can access the ***relevance tree***; and when exploring evaluation methodologies, they can reference the ***diff-table***

[1] https://rflow.ai/dashboard

for various evaluation approaches. For research activities requiring complex logical reasoning, the system would need additional guidance features to provide adequate support. While researchers could potentially leverage the system to support activities like module design and academic writing, such extensions would require integrating information sources beyond body-text of paper. For example, in system design, researchers need to reference figures from previous studies for detailed modeling. Similarly, during paper writing, our Research Survey Assistant Interface could add functionality to analyze deeper meta-information to help novice researchers organize references in each section, improving writing efficiency.

9.3.6 Research survey skill improvement

Our system focuses on helping researchers efficiently summarize large amounts of papers rather than improving their research skills. From an educational perspective, extending our system to develop researchers' top-down research skills could be a future direction. For example, we could implement a question-answering feature to enhance human-computer interaction. This would provide step-by-step guidance through top-down mind maps across multiple research directions based on novice researchers' questions, offering real-time feedback. Novice researchers could also modify mind map modules according to their preferences and ask follow-up questions about specific modules. This top-down question-answering approach could help novice researchers not only efficiently grasp research overviews but also strengthen their research survey skills.

Appendix A

UI

A.1 *Fish-bone*

A.2 *Relevance-tree*

A.3 *Diff-table*

Figure A.1: UI of *fish-bone*

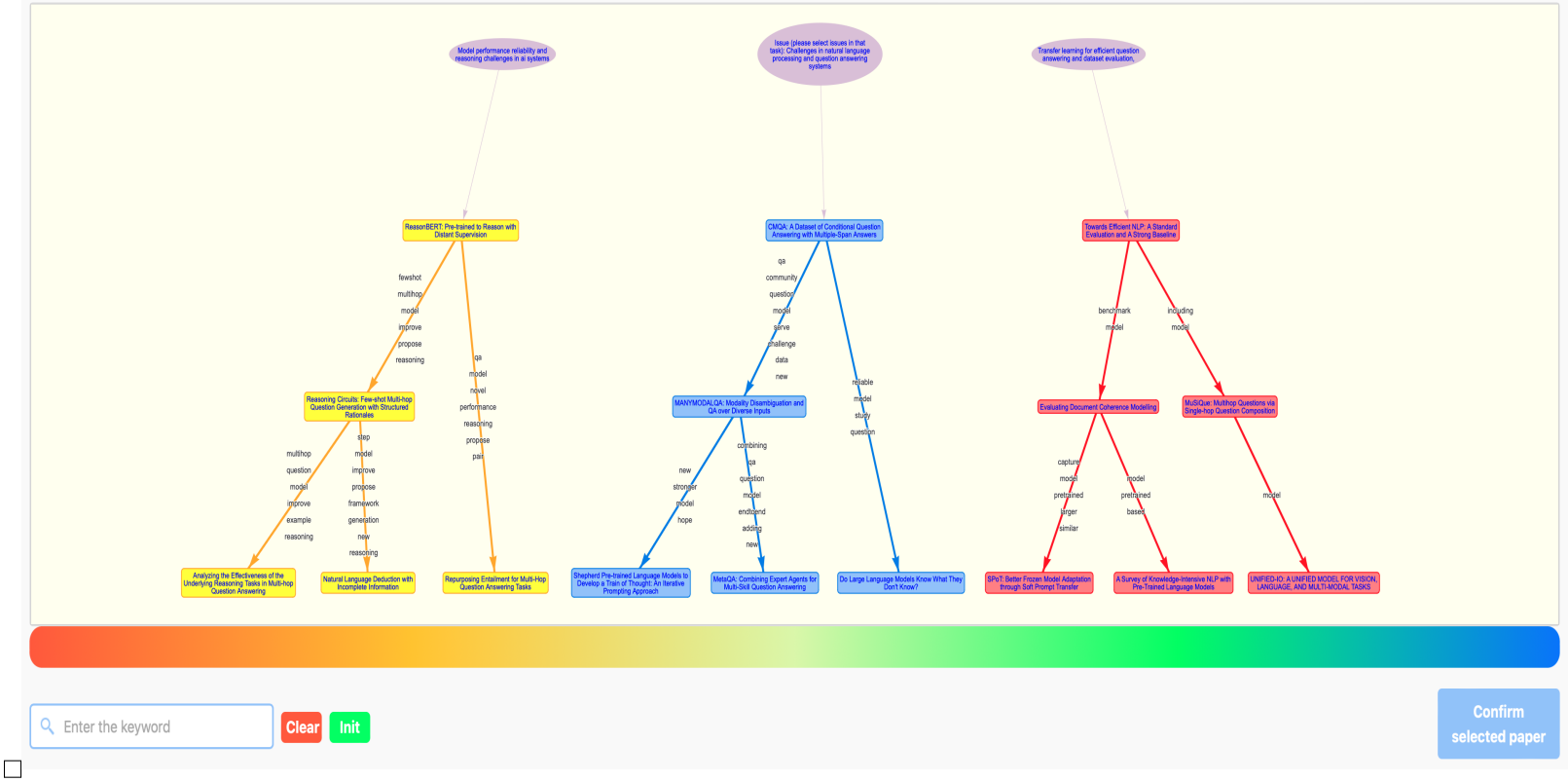


Figure A.2: UI of *Relevance-tree*

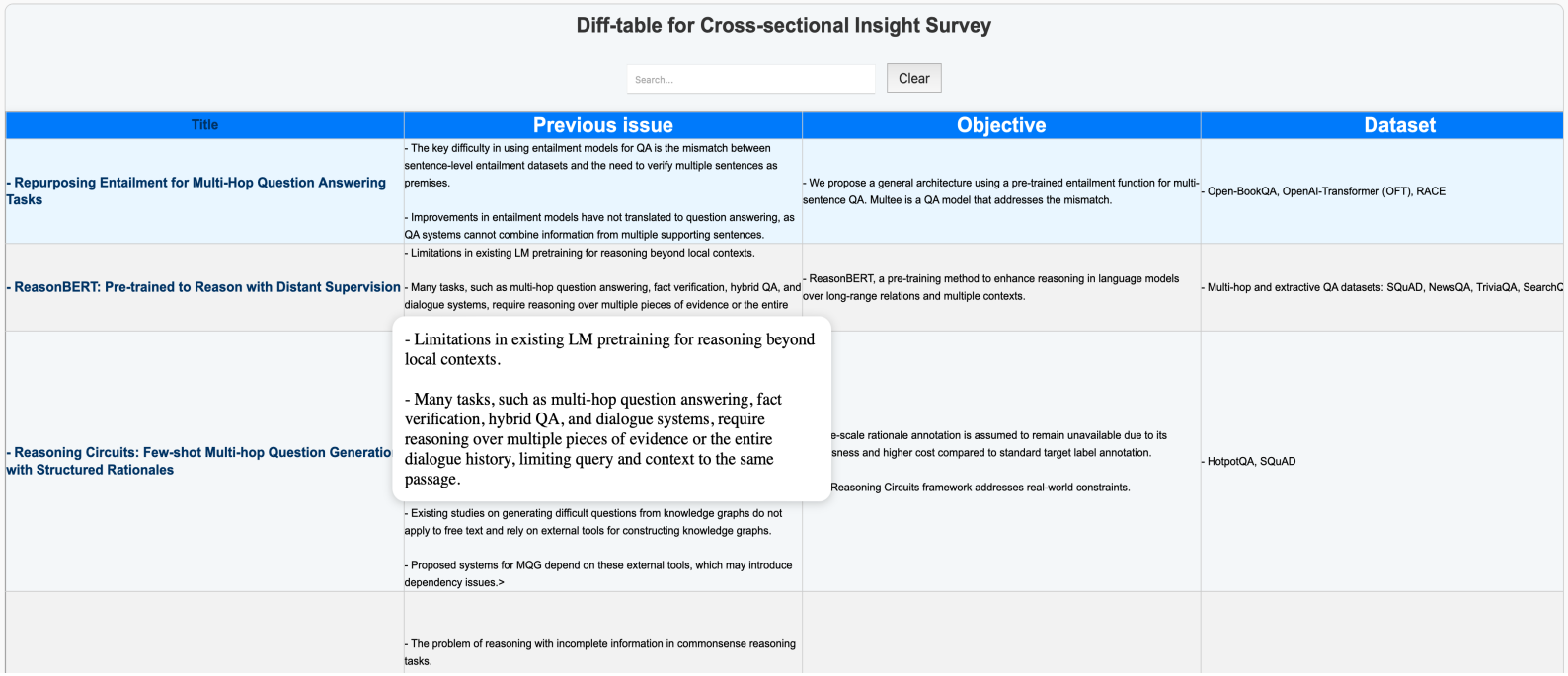


Figure A.3: UI of ***Diff-table***

Appendix B

Code of outcome evaluation

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Create DataFrame
data = {
    'No.': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11],
    'G': [1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4],
    'FAM': [1, 0, 0, 0, 0,
            0, 1, 1, 1,
            0, 0],
    'FRE': [2, 2, 1, 3, 2, 1,
            3, 4, 4, 1, 3],
    'APT': [3, 2, 3, 2, 2, 1, 3,
            1, 3, 2, 3],
    'Sufficient': [3, 2, 2, 1, 3, 1, 2, 2, 3, 2, 2],
    'Reasonableness': [2, 2, 3, 1, 1, 2, 2, 2, 2, 1, 2]
}
df = pd.DataFrame(data)

# Correlation analysis for numerical variables
correlation1 = df[['G', 'Sufficient', 'Reasonableness']].corr
    (method='spearman')
correlation2 = df[['FAM', 'Sufficient', 'Reasonableness']].
    corr(method='spearman')
correlation3 = df[['FRE', 'Sufficient', 'Reasonableness']].
    corr(method='spearman')
correlation4 = df[['APT', 'Sufficient', 'Reasonableness']].
    corr(method='spearman')

print(correlation1)
print(correlation2)
print(correlation3)
print(correlation4)
```

Listing B.1: Code of outcome evaluation

References

[1] Chand, D., & Ogul, H. (2020, March). Content-based search in lecture video: a systematic literature review. In 2020 3rd International Conference on Information and Computer Technologies (ICICT) (pp. 169-176). IEEE.

[2] Andersson, C., & Sundin, O. (2024). The elusive search engine: How search engine use is reflected in survey reports. Journal of the Association for Information Science and Technology, 75(5), 613-624.

[3] Cummings, J. (2021). Online navigation to journal articles: How are journal articles retrieved by researchers and students at an academic institution; a quantitative examination of HTTP referer [sic] data. Journal of Electronic Resources Librarianship, 33(2), 63-74.

[4] Lavidas, K., Achriani, A., Athanassopoulos, S., Messinis, I., & Kotsiantis, S. (2020). University students' intention to use search engines for research purposes: A structural equation modeling approach. Education and Information Technologies, 25, 2463-2479.

[5] Lhoest, Q., del Moral, A. V., Jernite, Y., Thakur, A., von Platen, P., Patil, S., ... & Wolf, T. (2021, November). Datasets: A Community Library for Natural Language Processing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 175-184).

[6] El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. Expert systems with applications, 165, 113679.

[7] Khan, B., Shah, Z. A., Usman, M., Khan, I., & Niazi, B. (2023). Exploring the landscape of automatic text summarization: a comprehensive survey. IEEE Access.

[8] Li, D., Qu, H., & Wang, J. (2023, November). A survey on knowledge graph-based recommender systems. In 2023 China Automation Congress (CAC) (pp. 2925-2930). IEEE.

[9] Rejeb, A., Rejeb, K., & Treiblmaier, H. (2023). Mapping metaverse research: identifying future research areas based on bibliometric and topic modeling techniques. Information, 14(7), 356.

[10] Libório, M. P., Martins, C. A. P., Laudares, S., & Ekel, P. I. (2023). Method of preparing an international and national literature review for novice researchers. MethodsX, 10, 102165.

[11] Gou, J., Sun, L., Yu, B., Wan, S., & Tao, D. (2023). Hierarchical multi-attention transfer for knowledge distillation. ACM Transactions on Multimedia Computing, Communications and Applications, 20(2), 1-20.

[12] Karunarathna, I., Gunasena, P., De Alvis, K., & Jayawardana, A. (2024). Structured reviews: Organizing, synthesizing, and analyzing scientific literature.

[13] Gao, J., Guo, Y., Lim, G., Zhang, T., Zhang, Z., Li, T. J. J., & Perrault, S. T. (2024, May). CollabCoder: a lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In Proceedings of the CHI Conference on Human Factors in Computing Systems (pp. 1-29).

[14] Sharifpour, R., Wu, M., & Zhang, X. (2023). Large-scale analysis of query logs to profile users for dataset search. Journal of Documentation, 79(1), 66-85.

[15] Kayyali, M. (2020). Post COVID-19: New era for higher education systems. International Journal of Applied Science and Engineering, 8(2), 131-145.

[16] Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020, October). SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 105-114).

[17] Phillips, E., & Johnson, C. (2022). How to Get a PhD: A handbook for students and their supervisors 7e. McGraw-Hill Education (UK).

[18] Hussain, T., Wang, D., & Li, B. (2024). The influence of the COVID-19 pandemic on the adoption and impact of AI ChatGPT: Challenges, applications, and ethical considerations. Acta Psychologica, 246, 104264.

[19] Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. Indian Journal of Ophthalmology, 71(12), 3600-3606.

[20] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017.

[21] 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.

[22] Lai, K. (2023). How well does ChatGPT handle reference inquiries? An analysis based on question types and question complexities. College & Research Libraries, 84(6), 974

[23] Hu, X., Tian, Y., Nagato, K., Nakao, M., & Liu, A. (2023). Opportunities and challenges of ChatGPT for design knowledge management. Procedia CIRP, 119, 21-28.

[24] Bian, N., Han, X., Sun, L., Lin, H., Lu, Y., He, B., ... & Dong, B. (2024, May). ChatGPT Is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 3098-3110).

[25] Rospigliosi, P. A. (2023). Artificial intelligence in teaching and learning: what questions should we ask of ChatGPT?. Interactive Learning Environments, 31(1), 1-3.

[26] Laban, P., Kryściński, W., Agarwal, D., Fabbri, A. R., Xiong, C., Joty, S., & Wu, C. S. (2023, December). SUMMEDITS: measuring LLM ability at factual reasoning through the lens of summarization. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 9662-9676).

[27] Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., ... & Nerdel, C. (2024). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. arXiv preprint arXiv:2401.00832.

[28] Ibrahim Altmami N, El Bachir Menai M. Automatic summarization of scientific articles: a survey. J King Saud Univ Comput InfSci. 2022;34(4):1011–28. https://doi.org/10.1016/j.jksuci.2020.04.020.

[29] G. Zaman, H. Mahdin, K. Hussain, Atta-Ur-Rahman, J. Abawajy and S. A. Mostafa, "An Ontological Framework for Information Extraction From Diverse Scientific Sources," in IEEE Access, vol. 9, pp. 42111-42124, 2021, doi: 10.1109/ACCESS.2021.3063181.

[30] Galal M. Binmakhashen and Sabri A. Mahmoud. 2019. Document Layout Analysis: A Comprehensive Survey. ACM Comput. Surv. 52, 6, Article 109 (November 2020), 36 pages. https://doi.org/10.1145/3355610

[31] Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. Information processing & management, 57(6), 102269.

[32] Ramakrishnan, C., Patnia, A., Hovy, E., & Burns, G. A. (2012). Layout-aware text extraction from full-text PDF of scientific articles. Source code for biology and medicine, 7, 1-10.

[33] Siegel, N., Lourie, N., Power, R., & Ammar, W. (2018, May). Extracting scientific figures with distantly supervised neural networks. In Proceedings of the 18th ACM/IEEE on joint conference on digital libraries (pp. 223-232).

[34] Da, C., Luo, C., Zheng, Q., & Yao, C. (2023, October). Vision Grid Transformer for Document Layout Analysis. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 19405-19415). IEEE.

[35] Lopez, P. (2009). GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds) Research and Advanced Technology for Digital Libraries. ECDL 2009. Lecture Notes in Computer Science, vol 5714. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04346-8_62.

[36] Clark, C., & Divvala, S. (2016, June). Pdffigures 2.0: Mining figures from research papers. In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (pp. 143-152).

[37] Frerebeau, N. (2019). tabula: an R package for analysis, seriation, and visualization of archaeological count data. Journal of Open Source Software, 4(44), 1821.

[38] Smock, B., Pesala, R., Abraham, R.: Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4634–4642(2022).

[39] Paliwal, S.S., D, V., Rahul, R., Sharma, M., Vig, L.: Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 128–133 (2019). https://doi.org/10.1109/ICDAR.2019.0002

[40] Xu, J., Kim, S., Song, M., Jeong, M., Kim, D., Kang, J., ... & Ding, Y. (2020). Building a PubMed knowledge graph. Scientific data, 7(1), 205.

[41] Zuluaga, M., Robledo, S., Arbelaez-Echeverri, O., Osorio-Zuluaga, G. A., & Duque-Méndez, N. (2022). Tree of Science - ToS: A Web-Based Tool for Scientific Literature Recommendation. Search Less, Research More!. Issues in Science and Technology Librarianship, (100). https://doi.org/10.29173/istl2696

[42] Chen, H., & Luo, X. (2019). An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing. Advanced Engineering Informatics, 42, 100959.

[43] Tu, Y., Qiu, R., & Shen, H. W. (2023). SKG: A Versatile Information Retrieval and Analysis Framework for Academic Papers with Semantic Knowledge Graphs. arXiv preprint arXiv:2306.04758.

[44] Chen, P. C., Huang, H. H., & Chen, H. H. (2022, November). 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) (pp. 384-389). IEEE.

[45] Duan, S., & Zhao, Y. (2023). Knowledge graph analysis of artificial intelligence application research in nursing field based on visualization technology. Alexandria Engineering Journal, 76, 651-667.

[46] Chen, W., Zhang, Y., Xian, Y., & Wen, Y. (2023). Hotspot Information Network and Domain Knowledge Graph Aggregation in Heterogeneous Network for Literature Recommendation. Applied Sciences, 13(2), 1093.

[47] Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: a survey. ACM Computing Surveys, 55(5), 1-37.

[48] da Silva, F. L., Slodkowski, B. K., da Silva, K. K. A., & Cazella, S. C. (2023). A systematic literature review on educational recommender systems for teaching and learning: research trends, limitations and opportunities. Education and Information Technologies, 28(3), 3289-3328.

[49] Kreutz, C. K., & Schenkel, R. (2022). Scientific paper recommendation systems: a literature review of recent publications. International journal on digital libraries, 23(4), 335-369.

[50] Pinedo, I., Larrañaga, M., & Arruarte, A. (2024). ArZiGo: A recommendation system for scientific articles. Information Systems, 122, 102367.

[51] Ahmedi, L., Rexhepi, E., & Bytyçi, E. (2021). Using association rule mining to enrich user profiles with research paper recommendation. International Journal Of Computing and Digital System.

[52] Chaudhuri, A., Sinhababu, N., Sarma, M. et al. Hidden features identification for designing an efficient research article recommendation system. Int J Digit Libr 22, 233–249 (2021). https://doi.org/10.1007/s00799-021-00301-2

[53] Hambarde, K. A., & Proenca, H. (2023). Information retrieval: recent advances and beyond. IEEE Access.

[54] Musaev, M., Rakhmatullaev, M., Normatov, S., Shukurov, K., & Abdullaeva, M. (2024, June). Integrated Intelligent System for Scientific and Educational Information Retrieval. In ENVIRONMENT. TECHNOLOGIES. RESOURCES. Proceedings of the International Scientific and Practical Conference (Vol. 2, pp. 212-219).

[55] Sharma, A., & Kumar, S. (2023). Machine learning and ontology-based novel semantic document indexing for information retrieval. Computers & Industrial Engineering, 176, 108940.

[56] McKeown, K., & Radev, D.R. (1995). Generating summaries of multiple news articles. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[57] Vassiliou, G., Papadakis, N., & Kondylakis, H. (2023, May). SummaryGPT: Leveraging ChatGPT for summarizing knowledge graphs. In European Semantic Web Conference (pp. 164-168). Cham: Springer Nature Switzerland.

[58] Zhang, R., Ouni, J., & Eger, S. (2024). Cross-lingual Cross-temporal Summarization: Dataset, Models, Evaluation. Computational Linguistics, 1-44.

[59] Li, J., Huy, P., Gu, W., Ota, K., & Hasegawa, S. (2024, September). Hierarchical Tree-structured Knowledge Graph For Academic Insight Survey. In 2024 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) (pp. 1-7). IEEE.

[60] Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., ... & Etzioni, O. (2018, June). 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) (pp. 84-91).

[61] Hayashi, H., Kryściński, W., McCann, B., Rajani, N., & Xiong, C. (2023, May). What's New? Summarizing Contributions in Scientific Literature. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 1019-1031).

[62] Liu, M. H., Yen, A. Z., Huang, H. H., & Chen, H. H. (2023, October). Contributionsum: Generating disentangled contributions for scientific papers. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 5351-5355).

[63] Chen, P. C., Huang, H. H., & Chen, H. H. (2022, November). 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) (pp. 384-389). IEEE.

[64] Faizullah, A. R. B. M., Urlana, A., & Mishra, R. (2024, August). LimGen: Probing the LLMs for Generating Suggestive Limitations of

Research Papers. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 106-124). Cham: Springer Nature Switzerland.

[65] Li, J., Gu, W., Ota, K., & Hasegawa, S. (2024). Object Recognition from Scientific Document Based on Compartment and Text Blocks Refinement Framework. SN Computer Science, 5(7), 816.

[66] Panda, Subhajit, Enhancing PDF Interaction for a More Engaging User Experience in Library: Introducing ChatPDF (June 23, 2023). IP Indian Journal of Library Science and Information Technology, 8(1), 20–25, 2023. https://doi.org/10.18231/j.ijlsit.2023.004.

[67] Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., ... & White, A. D. (2024). Language agents achieve superhuman synthesis of scientific knowledge. arXiv preprint arXiv:2409.13740.

[68] Block, J., & Kuckertz, A. (2024). What is the future of human-generated systematic literature reviews in an age of artificial intelligence?. Management Review Quarterly, 1-6.

[69] Whitfield, S., & Hofmann, M. A. (2023). Elicit: AI literature review research assistant. Public Services Quarterly, 19(3), 201–207. https://doi.org/10.1080/15228959.2023.2224125

[70] Sharma, R., Gulati, S., Kaur, A., Sinhababu, A., & Chakravarty, R. (2022). Research discovery and visualization using ResearchRabbit: A use case of AI in libraries. COLLNET Journal of Scientometrics and Information Management, 16(2), 215–237. https://doi.org/10.1080/09737766.2022.2106167

[71] Kaur, A., Gulati, S., Sharma, R., Sinhababu, A., & Chakravarty, R. (2022). Visual citation navigation of open education resources using Litmaps. Library Hi Tech News, 39(5), 7-11.

[72] Machado, L.M., Almeida, M.B., & Souza, R.R. (2020). What Researchers are Currently Saying about Ontologies: A Review of Recent Web of Science Articles. KNOWLEDGE ORGANIZATION.

[73] Michie, S., Hastings, J., Johnston, M., Hankonen, N.E., Wright, A.J., & West, R. (2022). Developing and using ontologies in behavioural science: addressing issues raised. Wellcome Open Research, 7.

[74] Li, J., Tanabe , H., Ota, K., Gu , W., & Hasegawa , S. (2023). Automatic Summarization for Academic Articles using Deep Learning and Reinforcement Learning with Viewpoints. The International FLAIRS Conference Proceedings, 36(1). https://doi.org/10.32473/flairs.36.133308

[75] Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company.

[76] Wang, X., & Cheng, Z. (2020). Cross-sectional studies: strengths, weaknesses, and recommendations. Chest, 158(1), S65-S71.

[77] Saier, T., Krause, J., & Farber, M. (2023, June). unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network. In 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 66-70). IEEE Computer Society.

[78] L. Jinghong, O. Koichi, G. Wen and H. Shinobu, "A Text Block Refinement Framework For Text Classification and Object Recognition From Academic Articles," 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), Hammamet, Tunisia, 2023, pp. 1-6, doi: 10.1109/INISTA59065.2023.10310320.

[79] Li, J., Gu, W., Ota, K. et al. Object Recognition from Scientific Document Based on Compartment and Text Blocks Refinement Framework. SN COMPUT. SCI. 5, 816 (2024). https://doi.org/10.1007/s42979-024-03130-7

[80] Weingart, P. (2017). The Future of Scholarly Publishing: Open Access and the Economics of Digital Publishing. African Minds.

[81] Ghosh, S., & Srivastava, S. (2022, May). ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3989-4004).

[82] Artifex: PyMuPDF 1.23.5 documentation (2015-2023). https://pymupdf.readthedocs.io/en/latest/

[83] Zhao, J., Zhang, T., Hu, J., Liu, Y., Jin, Q., Wang, X., & Li, H. (2022, May). M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5699-5710).

[84] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.

[85] Vasilakes, J., Zerva, C., Miwa, M., & Ananiadou, S. (2022, May). Learning Disentangled Representations of Negation and Uncertainty. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 8380-8397).

[86] Sollaci LB, Pereira MG. The introduction,methods,results,and discussion (imrad)structure: a fifty-year survey. J Med LibrAssoc. 2004;92(3):364.

[87] Syarif, I., Prügel-Bennett, A., Wills, G.B. (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. TELKOMNIKA Telecommunication Computing Electronics and Control, 14, 1502-1509.

[88] Muslim, M. A. (2020). Support vector machine (svm) optimization using grid search and unigram to improve e-commerce review accuracy. Journal of Soft Computing Exploration, 1(1), 8-15.

[89] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition, 30(7), 1145-1159.

[90] Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What Makes Reading Comprehension Questions Difficult?. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6951–6971, Dublin, Ireland. Association for Computational Linguistics.

[91] Cassidy, L., Lynn, T., Barry, J., & Foster, J. (2022, May). TwittIrish: a universal dependencies treebank of Tweets in modern Irish. Association for Computational Linguistics (ACL).

[92] Gan, L., Meng, Y., Kuang, K., Sun, X., Fan, C., Wu, F., & Li, J. (2022, May). Dependency Parsing as MRC-based Span-Span Prediction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2427-2437).

[93] Jie, Z., Li, J., & Lu, W. (2022, May). Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction.

In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5944-5955).

[94] Sugimoto, T., & Yanaka, H. (2022, May). Compositional Semantics and Inference System for Temporal Order based on Japanese CCG. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (pp. 104-114).

[95] Conforti, C., Berndt, J., Pilehvar, M. T., Giannitsarou, C., Toxvaerd, F., & Collier, N. (2022, May). Incorporating stock market signals for Twitter stance detection. In Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers) (pp. 4074-4091).

[96] Bikaun, T., Stewart, M., & Liu, W. (2022, May). Quickgraph: A rapid annotation tool for knowledge graph extraction from technical text. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 270-278).

[97] Ghosh, S., & Srivastava, S. (2022, May). ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3989-4004).

[98] Li, J., Shang, J., & McAuley, J. (2022, May). UCTopic: Unsupervised Contrastive Learning for Phrase Representations and Topic Mining. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6159-6169).

[99] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 2369-2380).

[100] Li, J., Phan, H., Gu, W., Ota, K., & Hasegawa, S. (2024). Fish-bone diagram of research issue: Gain a bird's-eye view on a specific research topic. *arXiv preprint arXiv:2407.01553*.

[101] Ahmad, R., Shaikh, Y., & Tanwani, S. (2023). Aspect Based Sentiment Analysis and Opinion Mining on Twitter Data Set Using Linguistic Rules. Indian Journal of Science and Technology, 16(10), 764-774.

[102] Reimers, N., & Gurevych, I. (2019, November). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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) (pp. 3982-3992).

[103] Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. TELKOMNIKA (Telecommunication Computing Electronics and Control), 14(4),1502-1509

[104] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

[105] Perrone, G., Unpingco, J., & Lu, H. M. (2020). Network visualizations with Pyvis and VisJS.

[106] Mavi, V., Jangra, A., & Jatowt, A. (2022). A survey on multi-hop question answering and generation. arXiv preprint arXiv:2204.09140.

[107] Joty, S., Carenini, G., Ng, R. T., & Murray, G. (2019). Discourse Analysis and Its Applications. https://doi.org/10.18653/v1/p19-4003

[108] Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019, August). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 319-327).

[109] Beltagy, I., Lo, K., & Cohan, A. (2019, November). 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) (pp. 3615-3620).

[110] Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. Information Processing & Management, 39(1), 45-65

[111] Ao, X., Yu, X., Liu, D., & Tian, H. (2020, June). News keywords extraction algorithm based on TextRank and classified TF-IDF. In 2020 International Wireless Communications and Mobile Computing (IWCMC) (pp. 1364-1369). IEEE.

[112] Shiu SH, Lehti-Shiu MD (2024) Assessing the evolution of research topics in a biological field using plant science as an example. PLOS Biology 22(5): e3002612. https://doi.org/10.1371/journal.pbio.3002612.

[113] Alto, V. (2023). Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4. Packt Publishing Ltd.

[114] Li, J., Phan, H., Gu, W., Ota, K., & Hasegawa, S. (2024). Fish-bone diagram of research issue: Gain a bird's-eye view on a specific research topic. arXiv preprint arXiv:2407.01553.

[115] Shadish, W.R., Cook, T.D., & Campbell, D.T. (2001). Experimental and Quasi-Experimental Designs for Generalized Causal Inference.

[116] Luo, Z., Xie, Q., & Ananiadou, S. (2023). ChatGPT as a Factual Inconsistency Evaluator for Abstractive Text Summarization. ArXiv, abs/2303.15621.

[117] Velásquez-Henao, J.D., Franco-Cardona, C.J., & Cadavid-Higuita, L. (2023). Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. DYNA.

[118] Dönmez, İ., Idin, S., & Gülen, S. (2023). Conducting Academic Research with the AI Interface ChatGPT: Challenges and Opportunities. Journal of STEAM Education, 6(2), 101-118. https://doi.org/10.55290/steam.1263404

[119] Rahman, Md. Mizanur and Terano, Harold Jan and Rahman, Md Nafizur and Salamzadeh, Aidin and Rahaman, Md. Saidur, ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples (March 27, 2023). Rahman , M., Terano, H. J. R., 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. doi: 10.52631/jemds.v3i1.175, Available at SSRN: https://ssrn.com/abstract=4407462

[120] Deng, C., Jia, Y., Xu, H., Zhang, C., Tang, J., Fu, L., ... & Zhou, C. (2021, October). GAKG: A multimodal geoscience academic knowledge graph. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 4445-4454).

[121] McKeown, K., & Radev, D.R. (1995). Generating summaries of multiple news articles. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

[122] Zhong, M., Liu, P., Chen, Y., Wang, D., Qiu, X., & Huang, X. J. (2020, July). Extractive Summarization as Text Matching. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6197-6208).

[123] Zhang, H., Zhang, X., Huang, H., & Yu, L. (2022, December). Prompt-based meta-learning for few-shot text classification. In Proceedings of the 2022 conference on empirical methods in natural language processing (pp. 1342-1357).

[124] Ma, K., Cheng, H., Zhang, Y., Liu, X., Nyberg, E., & Gao, J. (2023, July). 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) (pp. 1599-1618).

[125] Qi, P., Lin, X., Mehr, L., Wang, Z., & Manning, C. D. (2019, November). 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) (pp. 2590-2602).

[126] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[127] Inoue, N., Trivedi, H., Sinha, S., Balasubramanian, N., & Inui, K. (2021, November). Summarize-then-Answer: Generating Concise Explanations for Multi-hop Reading Comprehension. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 6064-6080).

[128] Aharoni, R., Narayan, S., Maynez, J., Herzig, J., Clark, E.,& Lapata, M. (2023, July). Multilingual Summarization with Factual Consistency Evaluation. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 3562-3591).

[129] Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, 391-409.

[130] Schober, Patrick MD, PhD, MMedStat; Boer, Christa PhD, MSc; Schwarte, Lothar A. MD, PhD, MBA. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 126(5):p 1763-1768, May 2018. — DOI: 10.1213/ANE.0000000000002864

[131] Lee, M. S. (2024). Improving the Transparency of Agent Decision Making to Humans Using Demonstrations (Doctoral dissertation, University of Massachusetts Amherst).

[132] Haensel, M., Schmitt, T. M., & Bogenreuther, J. (2023). Teaching the Modeling of Human–Environment Systems: Acknowledging Complexity with an Agent-Based Model. Journal of Science Education and Technology, 32(2), 256-266.

[133] Zhang, W., Sjoerds, Z., & Hommel, B. (2020). Metacontrol of human creativity: The neurocognitive mechanisms of convergent and divergent thinking. NeuroImage, 210, 116572.

[134] Yan, Y. H., & Chien, T. W. (2021). The use of forest plot to identify article similarity and differences in characteristics between journals using medical subject headings terms: a protocol for bibliometric study. Medicine, 100(6), e24610.

[135] Aljuaid, H., Iftikhar, R., Ahmad, S., Asif, M., & Afzal, M. T. (2021). Important citation identification using sentiment analysis of in-text citations. Telematics and Informatics, 56, 101492.

[136] Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2023). A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations. Journal of Information Science, 49(1), 79-92.

Publication

International journal

1. Li, J., Gu, W., Ota, K., Hasegawa, S., Object Recognition from Scientific Document Based on Compartment and Text Blocks Refinement Framework. SN COMPUT. SCI. 5, 816 (2024). https://doi.org/10.1007/s42979-024-03130-7

International conferences (peer review)

1. Jinghong Li, Naoya Inoue, Shinobu Hasegawa, A Viewpoints Embedded diff-table System for Cross-sectional Insight Surveys In a Research Task (The 38th Pacific Asia Conference on Language, Information and Computation in press)

2. J. Li, H. Phan, W. Gu, K. Ota and S. Hasegawa, "Fish-Bone Diagram of Research Issue: Gain a Bird's-Eye View on a Specific Research Topic," 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 2024, pp. 4936-4941, doi: 10.1109/SMC54092.2024.10830995.

3. J. Li, W. Gu, K. Ota and S. Hasegawa, "A Survey Forest Diagram: Gain a Divergent Insight View on a Specific Research Topic," 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Kuching, Malaysia, 2024, pp. 4136-4137, doi: 10.1109/SMC54092.2024.10831785.

4. J. Li, P. Siritanawan, W. Gu and S. Hasegawa, "Multi-Agent Approach for Dynamic Research Insight Path Generation," 2024 IEEE International Conference on Agents (ICA), Wollongong, Australia, 2024, pp. 128-129, doi: 10.1109/ICA63002.2024.00035.

5. J. Li, P. Huy, W. Gu, K. Ota and S. Hasegawa, "Hierarchical Tree-structured Knowledge Graph For Academic Insight Survey," 2024 International Conference on INnovations in Intelligent SysTems and Appli-

cations (INISTA), Craiova, Romania, 2024, pp. 1-7, doi: 10.1109/INISTA62901.2024.10683856.

6. L. Jinghong, O. Koichi, G. Wen and H. Shinobu, "A Text Block Refinement Framework For Text Classification and Object Recognition From Academic Articles," 2023 International Conference on Innovations in Intelligent Systems and Applications (INISTA), 2023, pp. 1-6, doi: 10.1109/INISTA59065.2023.10310320.

7. Jinghong Li, Hatsuhiko Tanabe, Koichi Ota, Wen Gu, and Shinobu Hasegawa: Automatic Summarization for Academic Articles using Deep Learning and Reinforcement Learning with Viewpoints. The International FLAIRS Conference Proceedings,36(1),2023. https://doi.org/10.32473/flairs.36.133308.

Preprint

1. Li, J., Gu, W., Ota, K., & Hasegawa, S. (2024). A Survey Forest Diagram: Gain a Divergent Insight View on a Specific Research Topic. arXiv preprint arXiv:2407.17081.

2. Li, J., Phan, H., Gu, W., Ota, K., & Hasegawa, S. (2024). Fish-bone diagram of research issue: Gain a bird's-eye view on a specific research topic. arXiv preprint arXiv:2407.01553.

3. Li, J., Siritanawan, P., Gu, W., & Hasegawa, S. Multi-Agent Approach for Dynamic Research Insight Path Generation.

4. Li, J., Inoue, N, & Hasegawa, S. A Viewpoints Embedded Diff table System For Cross sectional Insight Survey In a Research Task.

5. Li, J., Phan, H., Gu, W., Ota, K., & Hasegawa, S. (2024). Hierarchical Tree-structured Knowledge Graph For Academic Insight Survey. arXiv preprint arXiv:2402.04854.

6. Li, J., Gu, W., Ota, K., & Hasegawa, S. (2023). Object Recognition from Scientific Document based on Compartment Refinement Framework. arXiv preprint arXiv:2312.09038.

Domestic conference

1. Hatsuhiko Tanabe, Jinghong Li, Shinobu Hasegawa: "Building a summary generation system that reflects the viewpoint of entering academic papers", Research Report of Japanese Society for Information and Systems in Education Vol.38(4), pp. 1-7, (2023)

2. Li Jinghong, Koichi Ota, Gu Wen, Shinobu Hasegawa: "Development of automatic data set construction system for text analysis tasks in academic literature", Research Report of Japanese Society for Information and Systems in Education Vol.37 (7), pp.39-46, (2023)

3. Li Jinghong, Koichi Ota, Shinobu Hasegawa: "Automatic summary generation of academic papers by deep learning and reinforcement learning reflecting the viewpoint" Research Report of Japanese Society for Information and Systems in Education, 36(1), pp.68-73, (2021)

4. Jinghong, L., Koichi, O., & Shinobu, H. (2020). Automatic Extracted Summary Generation with Viewpoints for Research Articles based on Deep Reinforcement Learning. IEICE Technical Report; IEICE Tech. Rep., 120(167), 53-56.

Dataset

1. LI JINGHONG, and Huy Phan. (2024). Dataset for academic insight survey [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/4330260.

Source code

1. Hierarchical Tree-structured Knowledge Graph:Gain an Insight View for Specific research Topic: https://github.com/Hasegawa-lab-JAIST/LI_JINGHONG_Hierarchical-Tree-structured-Knowledge-Graph-For-Academic-Insight-Survey