Paper Title* (use style: paper title)

*Note: Sub-titles are not captured in Xplore and should not be used

line 1: 1st Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 2nd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 3rd Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 4th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 5th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

line 1: 6th Given Name Surname
line 2: *dept. name of organization (of Affiliation)*
line 3: *name of organization (of Affiliation)*
line 4: City, Country
line 5: email address or ORCID

***Abstract*—The exponential growth of academic articles on the Internet due to expanding research domains has rendered the retrieval of target articles a notably time-consuming endeavor. A majority of these scholarly publications presented in the form of PDFs, a factor that sadly contributes to substantial information loss, especially mathematical expressions. In response, we introduce "arxiv summarizer", a novel Scientific Automatic Summarization System (SASS) based on Nougat , converting document pages to markup , and Large Language Models(LLMs) like GPT-3.5, to create concise and easily digestible summaries. The proposed solution facilitates the interactive generation of precise and fluent summaries suitable for academic papers, interpretive blogs, and broadcasts. Our experimental results showcase the effectiveness of SASS, demonstrating its capability to outperform even GPT-4. This positions it as a reliable option to expedite the retrieval of target documents in the scholarly landscape.**

<!-- TODO: Figure -->

***Keywords*—component, formatting, style, styling, insert (*key words*)**

I. Introduction (*Heading 1*)

The advent of the digital era has precipitated a swift escalation in the online dissemination of academic articles, rendering the acquisition of scientific documents of interest an increasingly formidable task for researchers. Query-based searches in certain scientific fields often yield a vast assortment of pertinent articles, the volumes of which surpassing human processing abilities [7]. This is primarily due to the fact that scientific literature tends to be voluminous in nature.

Various system applications geared towards condensing long documents have been proposed by researchers. These practices include writing section-structured [82], user-oriented [45], or presentation-specific [107] summaries for scientific papers, and the automation of scientific reviews [126] aiming to distill protracted documents into their essential constituents, presented in multiple forms.

While these strategies have made considerable strides towards the objective of goal of long document summarization, their applicability in summarizing scientific documents for user to filter what they interested remains limited. The reasons for this are multi-faceted, involving areas where information is most likely to be lost during processing: 1) Predominantly, the methodologies mentioned above are designed to process data presented in plain text format, leaving them incapable of processing documents in PDF format. This limitation poses significant hindrances to users. 2) A significant portion of prior approaches overlook the multimodal data available in scientific documents, such as mathematical formulas, tables, figures, etc. which encapsulate the most crucial experimental results, concepts, or workflows[6]. 3) Lastly, many strategies are either incapable in handling long documents, as they often prioritize short document practice or merely employ a simplistic aggregation of extracted sentences, which is hard for user to follow.

In light of the ongoing issues, i.e. the challenges posed by PDF document parsing, capitalizing on essential multimedia information, and summarization of lengthy texts, we propose "arXiv Summarizer", an innovative Scientific Automatic Summarization System (SASS) which offers a resolution to these dilemmas. "arXiv Summarizer" provides clear, engaging, and easily comprehensible summaries that include original text figures, tables, and key mathematical formulas as an aid to understanding. The produced summary allows for a comprehensive understanding of the principles and core content even without the necessity of reading the original text. Furthermore, the results can be readily transformed into various formats (e.g., blogs, broadcasts, etc.).

Our system draws upon the latest developments in Optical Understanding to address the first two difficulties. When it comes to summarizing long articles, the content generated by extractive-based mechanisms is essentially a mere compilation of raw sentences derived from the document. In real-world settings, however, summaries frequently resort to abstractive summarization. However, a significant limitation of these abstractive methods is their capacity constraints, often limited to processing between 512 to 1024 lexical tokens [26-27], due to the memory complexity issues and hardware limitations and the time costing pretraining process. In response to this, we employ a hybrid approach that utilizes a generally hierarchical discourse structure to efficiently divide the long document into shorter, section-level sub-articles. Subsequently, we identify important fragments and integrate these into the article-level interpretation using abstractive methods. Given that the content selection mechanism is one of two most notable long document mechanism[40], and considering the inflexibility of fine-tuning pre-trained models in the summarization domain, we opted for the efficient, prompt-based Large Language Models (LLMs) as an alternative.

The core contributions advanced by our paper are:

1. We introduce a dynamic, interactive system capable of accepting PDF inputs while generating concise and lucid summaries equipped with multimedia information. These summaries can be further modified and utilized for academic papers, interpretative blogs, and broadcasts, thereby bridging the gap between academic research papers in PDF format, and those gold summaries carefully crafted by human experts in summarizations of scientific articles.

2. We propose a pipeline for processing long scientific text into distinct groups and subsequently merging them into lucid, seamless summaries.

II. RELATED WORK

PDF Document Parsing Most of scientific documents are in the Portable Document Format (PDF), which comprises 2.4% of the common crawl on the internet, making it the second most prevalent data format online [1]. Despite this ubiquity, extracting information embedded in these academic resources into other formats poses a significant challenge, particularly for specialized documents like scientific research papers. This complexity is further compounded when dealing with mathematical expressions, whose semantic information is often lost [2].

Past technologies such as Optical Character Recognition (OCR) engines, exemplified by Tesseract OCR [3], have proven effective in extracting individual characters and words from images, even in intricate environments. Nonetheless, these engines fall short in preserving the relative positional relationship among different formats, especially mathematical expressions and tables due to their line-by-line approach. This limitation underscores a severe shortfall in information capture, hampering the generation of accurate summaries even for human readers who might misconstrue the intended meanings without the original context.

Text Summarization The growth in summarizing generic texts since the 1950s has been commendable [4]. Yet, summarizing scientific articles presents a unique challenge, despite they have a more structured nature, generally comprising sections like introduction, methodology, experiments, and conclusions. As pointed out by Teufel and Moens (2002) [5], there are two main reasons that make scientific articles notably different: 1) scientific articles are typically lengthier than generic texts, and 2) the objectives of the summary are multifaceted, as various researchers may be interested in different aspects, such as new contributions, findings, and solutions proposed.

Bhatia and Mitra (2012) [6] reinforced another issue, positing that conventional text methodologies fall short in creating summaries for academic research due to most of the information value embodied in 'document elements. These elements, separate from the running text, either augment or summarize the text's content.Figures, tables, mathematical expressions, and pseudocodes for algorithms often constitute these elements, representing the most pertinent experimental results, ideas, or workflows.

The abstract is a plausible substitute for a summary. However, several issues arise: 1) Abstracts typically do not delve into the specifics of the full text [7], 2) they frequently represent the author's subjective and potentially biased viewpoint of distinctive features [8], 3) a single abstract does not necessarily meet every user's requirements [9], and 4) an abstract may not encompass all the paper's impacts and contributions but highlights what the author desires to emphasize (Elkiss et al., 2008) [10]. As such, the ideal summary from a Scientific Automatic Summarization System should capture the essence and maintain objectivity while being revealing.

Current prevailing methods for scientific article summarization are primarily categorized into two classes:abstract generation-based approaches and citation-oriented approaches. Their classification is contingent upon whether the abstract is considered the reference summary or we called gold summary.

The abstract generation-based approaches aim to automatically generate an abstract for a research article, implying that the abstract is deemed the gold summary, with the remaining text utilised as input. Noteworthy research in this area includes the work of Yang et al. (2016)[11], who designed a system for generating an extended abstract that describes the most important aspects of a scientific article employing a data-weighted reconstruction approach. This process comprises two stages: weight learning and salient sentence selection. Similarly, Slamet et al. (2018)[12] proposed a system to generate article abstracts automatically for the Indonesian language. Both employ intricate mechanisms for sentence weight calculation. The finalized abstract, compiled from the top ranked sentences, may subsequently undergo a reconstruction process.

Following the advent of BERT [13], numerous large-scale models, each with distinct pre-training tasks, have been introduced. However, due to hardware constraints, the input token length is limited to 512 tokens. To address this limitation, Cui and Hu [14] proposed a memory network that incorporates graph attention networks and gated recurrent units to dynamically select important sentences through sliding a window along the entire source document. By confining its use within each window—where the window size is set to be less than or equal to 512 tokens—this approach effectively employs the pre-trained BERT model to tackle long document summarization tasks.

Contrastingly, citation-oriented approaches operate under the assumption that citation sentences typically encapsulate the essential information from the cited article. Thus, citation sentences towards a target paper could be viewed as a succinct summary penned by the cited researchers. Citation-oriented approaches can potentially outperformabstract generation-based approaches as citation summaries reflect the viewpoints of multiple researchers, while the abstract signifies only the authors' perspectives.

In the landscape of citation-oriented methodologies, the crux of most proposals centers on sentence ranking, augmentation, classification, and summarization. This is exemplified in the work of Lauscher et al. (2017)[15] who developed a system leveraging their Learning to Rank (L2R) model. This model ranks sentences from the target paper based on their similarity. Following this initial stage, they engage in clustering and sentence selection, guided by the TextRank score (Mihalcea and Tarau, 2004)[16].Furthermore, Agrawal et al. (2019)[17] proposed a semi-supervised method to tackle scientific article summarization.

As summarization with the abstractive approach is naturally a sequence-to-sequencetask,where the source and target reside in different spaces due to variations in length, redundancy, and other metrics. Thus, a pre-trained model with a sequence-to-sequence objective task would be more fitting than using an encoder-only (e.g., BERT/RoBERTa) or a decoder-only (e.g., GPT-2/GPT-3) pre-trained model. Within the domain of summarization, Bidirectional and Auto-Regressive Transformers (BART) [18] and Text-to-Text Transfer Transformers (T5) [19] are the two preeminent sequence-to-sequence pre-trained models.

BART is pre-trained on a self-supervised task of reconstructing arbitrarily corrupted text while T5 is pre-trained on both unsupervised and supervised objectives, such as token masking, as well as translation and summarization.

Despite the demonstrable effectiveness of T5 in the domain of short document summarization [20], no supervised transformer models within the long document summarization domain - let alone the scientific article domain - have incorporated a summarizer equipped with a T5 pre-training task. While, BART, pre-trained on short documents, adheres to an input limitation of 1,024 tokens.

Summarily, the aforementioned methods can reasonably be characterized as models that train on extensive summarization datasets tailored to specific domains. However, these models require fine-tuning when transitioning to unfamiliar data distributions. Notably,most of them are extractive approach which often extract key sentences from the original text and integrate them. The prevalence of extractive approaches can be attributed to scientific papers falling under the category of domain-specific articles, which tend to feature more complex formulas and terminologies. Nonetheless, this indicates that a model that merely extracts lexical fragments from the original text of a long document can still generate a summary closer in resemblance to the reference summary. Given that abstractive summarization models have recently been found to contain factual inconsistencies in up to 30% of the summary outputs in the short document domain [21-22] while extractive summarization model will faithfully preserve the original content, this insight encourages for the development of long document extractive models in the real-world production level settings.However, extractive approaches are inadequate to summarize a long scientific article into a succinct and fluent summary, particularly when the salient contents are thinly scattered[23] and may overlap with each other, despite may subsequent attempts at reconstruction.

The recent success of large language models, such as GPT-3, has prompted a paradigm shift in Natural Language Processing research. Large language models (LLMs) provide a novel methodology for generating summaries. They leverage natural language task instructions and exhibit zero or few-shot performances in the prompt without the requirement for weight updates. LLMs demonstrate overwhelming preference by humans when compared to their fine-tuning pretrained model counterparts and remain unaffected by common dataset-specific problems like insufficient factual accuracy in news summarization.[24]and cross-lingual summarization[25].

Long Document Summarization Recent Long Scientific Document models are mainly developed on the scientific paper benchmark dataset, arXiv/PubMed. There are some relevant systems aiming in the application for research-related tasks. For instance, CTRLSUM[45], aiming in crafting user-specific summaries, introduces an innovative framework for controllable summarization. This involves the deployment of control tokens designed as a set of keywords or descriptive prompts, and a BERT-based sequence tagger computes the selection probability for each token in the test document. Likewise, D2S[107], crafting presentation-oriented summaries, incorporate a Dense Vector IR module to pinpoint the most relevant sections/sentences, alongside figures/tables from the related paper. With a QA model, this generates an abstractive summary (answer) of the retrieved text, providing a solution for the document-to-slides.

Other related work is discussing the reliability of automatically generated abstracts for scientific papers [126] and generating literature survey based on multiple biomedical long scientific papers [24]. Models designed for long document summarization also have utility across additional domains. They can be employed for auxiliary tasks such as video captioning [72], question-answering from lengthy documents [76] or multi-modal tasks [68, 86].