MASSW: A New Dataset and Benchmark Tasks for AI-Assisted Scientific Workflows

Xingjian Zhang[1*] **Yutong Xie**[1*] **Jin Huang**[1] **Jinge Ma**[2] **Zhaoying Pan**[2]
Qijia Liu[1] **Ziyang Xiong**[1] **Tolga Ergen**[3] **Dongsub Shim**[3]
Honglak Lee[1,3] **Qiaozhu Mei**[1]

[1]University of Michigan, Ann Arbor [2]Purdue University [3]LG AI Research
[1]{jimmyzxj,yutxie,huangjin,ponypony,xziyang,qmei}@umich.edu honglak@eecs.umich.edu
[2]{ma859, pan433}@purdue.edu
[3]{tergen, dongsub.shim}@lgresearch.ai

Abstract

Scientific innovation relies on detailed workflows, which include critical steps such as contextualizing literature, generating ideas, validating ideas, interpreting results, and planning new research. Scientific publications that document these workflows are extensive and unstructured, making it difficult to effectively navigate and explore the space of scientific innovation. To meet this challenge, we introduce **MASSW**, a comprehensive dataset of **M**ulti-**A**spect **S**ummarization of **S**cientific **W**orkflows. MASSW includes more than 152,000 peer-reviewed publications from 17 leading computer science conferences spanning the past 50 years. Using Large Language Models (LLMs), we automatically extract five core aspects from these publications – *context*, *key idea*, *method*, *outcome*, and *projected impact* – which correspond to five key steps in a research workflow. We show that these LLM-extract summaries have a comparable quality to human annotations, and they facilitate a variety of downstream tasks, corresponding to different types of predictions and recommendations along the scientific workflow. Overall, MASSW demonstrates decent utility as a pre-computed and trustful resource for the AI4Science community to create and benchmark a wide-range of new AI methods for optimizing scientific workflows and fostering scientific innovation. Our dataset is available at `https://huggingface.co/datasets/jimmyzxj/massw`.

1 Introduction

Can AI be a capable copilot for scientific research? Scientific innovation is driven by complex and detailed workflows, also referred to as scientific methods at a coarse level (Ayala, 2009; Voit, 2019). These workflows typically involve critical steps such as analyzing existing literature, generating research ideas, validating these ideas through analyses and experiments, interpreting the results, and ultimately inspiring new research inquiries. To navigate and explore the space of innovations, both the pilot and the copilot have to understand, plan, and optimize the scientific workflows (Wang et al., 2023). These workflows are widely documented in scientific publications, serving as a key source for scientists to understand, reproduce, and plan research. However, these publications are typically unstructured and complex, making it difficult to trace scientific workflows or extend them to new research. To support researchers, or even AI copilots, in better navigating and exploring the scientific innovation landscape, it is crucial to develop new datasets that document scientific workflows in a more structured and "ready-to-analyze" manner, along with new tools that enable reasoning and evolution of these workflows.

Curating scientific workflow datasets from scientific publications is challenging. While human experts are skilled at deciphering complex scientific publications, their highly personalized interpretations, if not sufficiently aligned, often result in inconsistent and heterogeneous annotations and predictions (Beck et al., 2020). Furthermore, annotations by highly specialized researchers are inherently expensive, limiting the feasibility of building large datasets at the scale and scope of a scientific field (Takeshita et al., 2024; Fisas et al., 2015; Cachola et al., 2020a; Mei and Zhai, 2008). These challenges highlight the need for an automated, scalable, and consistent solution to annotate structured scientific workflows, a task well-suited for an AI. Indeed, recent large language models (LLMs) have demonstrated promising performance in reasoning through natural language (Wei et al., 2024), positioning them as a viable candidate for automating the annotation of scientific workflows, even though it remains to be seen whether they can

*Equal contribution

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2373–2394

match the accuracy of human experts.

To address these challenges, we present **MASSW**, a novel and large-scale dataset that provides a comprehensive and structured **M**ulti-**A**spect **S**ummarization of **S**cientific **W**orkflows. The key features of MASSW include

- **Structured scientific workflows**. MASSW defines five core aspects of a scientific workflow – *context*, *key idea*, *method*, *outcome*, and *projected impact*. These aspects correspond to the typical stages of general scientific workflows found in the literature. Utilizing LLMs, we are able to consistently extract and structure these five aspects from each publication.
- **Large scale**. MASSW contains the structured scientific workflows and meta-information from over 152,000 peer-reviewed publications, across 17 leading computer science conferences, and spanning the past 50 years.
- **Accuracy**. The coverage and accuracy of MASSW have been validated through comprehensive inspections and comparisons with expert annotations and alternative approaches.

MASSW provides a resource to build a wide range of applications in AI-assisted scientific discovery, offering a structured, large-scale, and precomputed data source. It can be used to benchmark various AI-driven scientific research challenges such as generating new scientific ideas, designing experiments, hypothesis validation, and forecasting the impact of research. For scientists, MASSW can be used to facilitate their exploratory analysis of the scientific space, possibly through visualizations of the literature and workflows. One can utilize it to quantify the novelty of ideas within a historical context (Wang et al., 2023). For AI researchers, this unique data source can also be used to develop search engines and recommender systems to retrieve similar ideas, suggest methods, and support fine-grained examination of scientific workflows (Hope et al., 2018). While this paper does not demonstrate all potential applications of MASSW, we benchmark several downstream tasks that can be evaluated without specific human judgments, in order to showcase MASSW's versatility and ability to open new opportunities for new AI applications and methods. Ultimately, we anticipate datasets like MASSW could foster more effective scientific workflows and accelerate innovation.

2 Dataset Overview

MASSW is a *structured*, *large-scale*, and *precomputed* dataset designed to enhance the exploration and analysis of scientific workflows. In Section 2.1, we first discuss how a scientific publication can be structured into five core aspects, corresponding to five key steps in a general scientific research workflow. In Section 2.2, we describe the curation of scientific publication data and an automated procedure that summarizes these core aspects with LLMs. Lastly, we present basic statistics about the constructed MASSW dataset and a multi-view visualization of these aspects in Section 2.3.

2.1 Core Aspects of Scientific Workflows

A typical scientific research workflow generally follows a series of steps: posing a research question, reviewing existing literature, formulating a hypothesis or research idea, validating the hypothesis, interpreting results, drawing conclusions, reporting findings, and planning future research (Ayala, 2009; Voit, 2019). In a scientific publication, these steps are often described through specific narrative elements. For instance, authors typically situate their study within the *context* of existing work, present the *key idea* driving their research, describe the *method* used to test the idea, discuss the *outcome* of their validation, and highlight the *potential impact* of their findings. In Table 1, we formally define these core aspects of a scientific workflow. While various studies have proposed their own definitions of related components, their focus has largely been on testing NLP models' ability to extract and summarize information from scientific publications (Fisas et al., 2015; Takeshita et al., 2024; Fok et al., 2023; Cohan et al., 2019; Dernoncourt and Lee, 2017), rather than establishing a large-scale dataset that documents essential aspects of scientific workflows. Inspired particularly by Fisas et al. (2015), whose categorization framework is theoretically grounded in the field of computer science, we adopt the following five core aspects of scientific workflows:

Context The context of a study summarizes the status quo of the research field or the broader reality before the study is published. This aspect is often related to analyzing relevant literature, identifying the gap and unresolved challenges, and motivating

Aspect	Definition	Example
Context *Ask questions, review literature (prior to study)*	The status quo of related literature or reality which motivated this study. This could normally be a problem, a research question, or a research gap that has not been successfully addressed by previous work.	*Making language models bigger does not inherently make them better at following a user's intent, as large models can generate outputs that are untruthful, toxic, or not helpful.*
Key Idea *Construct hypothesis (proposed in this study)*	The main intellectual merit of this paper, often in comparison to the context. This could normally be a novel idea or solution proposed in this paper that distincts it from what's already done in literature.	*The authors propose InstructGPT, a method to align language models with user intent by fine-tuning GPT-3 using a combination of supervised learning with labeler demonstrations and reinforcement learning from human feedback.*
Method *Test hypothesis (after hypothesis construction)*	The specific research method that investigates and validates the key idea. This could be an experimental setup, a theoretical framework, or other necessary validation methodology to implement and/or evaluate the key idea.	*The authors evaluate the performance of InstructGPT by humans on a given prompt distribution and compare it with a much larger model GPT-3.*
Outcome *Interpret results, draw conclusion (after testing hypothesis)*	The factual statement about the study output. This could be the experiment results and any other measurable outcome that has occurred. It marks whether the key hypothesis is testified or not.	*InstructGPT, even with 100x fewer parameters, is preferred over GPT-3 in human evaluations. It shows improvements in truthfulness and reductions in toxic outputs with minimal performance regressions on public NLP datasets.*
Projected Impact *Future work (anticipated but not yet done)*	The author-anticipated impact of the work on the field, and potential further research identified by the author that may improve or extend this study.	*Fine-tuning with human feedback is a promising direction for aligning language models with human intent.*

Table 1: Core aspects in the MASSW dataset that correspond to key steps (*in italic*) in a general scientific workflow. The example is based on the paper "Training Language Models to Follow Instructions with Human Feedback." (Ouyang et al., 2022) More examples of MASSW are provided in Appendix B.

new research ideas to fill the gap. In particular publications, this key aspect is often described as *background*, *challenges*, or *literature review*, as adopted by previous work of text summarization.

Key Idea The key idea represents the central hypothesis or novel contribution proposed in the study. This is the key aspect that distinguishes the current work from the context of existing work. It is a product of idea generation, a critical step in the scientific workflow where new concepts are formed, new connections are made, and new solutions are proposed to address particular challenges in research. In previous work of text summarization, it is sometimes related to the *approach* described in a paper, which only partially reflects its key ideas.

Method The method of a study details the procedures and techniques used to validate the key idea or hypothesis. In other words, the method is not a part of the hypothesis itself, but rather the procedure used to prove or reject the hypothesis. In previous work of text summarization, *method* is sometimes confused with the *key idea* (both referred to as part of the *approach* (Fisas et al., 2015; Takeshita et al., 2024)), especially when the main subject of the research is a "method." We explicitly distinguish *method* from the *key idea* as they refer to different steps in the scientific workflow (generating ideas v.s. validating ideas).

Outcome The outcome includes the results and findings as a product of the *method* in the study. This aspect corresponds to the measurable results, the interpretation of these results, and other types of impact of the work that has already happened by the time of publication. This concept is also mentioned in previous work of text summarization, as *outcome* or *result*.

Projected Impact The projected impact outlines the potential future implications of the research that have not happened at the time of publication. This aspect is often an ex-ante prediction of how the results of the work would inspire follow-up research or deployment, from the author's point of view. It discusses how the findings can contribute to the field, suggests new research directions, and potentially leads to societal or technological advancements. Previous text summarization work often simply uses the concept of *future work*, while ignoring the broader impact of the study.

2.2 Data Curation and Aspect Summarization

Advancing AI in understanding and improving scientific workflows requires large-scale and high-quality data. To address this challenge, we curate a collection of scientific publications and structure it into the above-defined core aspects at scale.

Large-scale scientific publication collection. To build this initial version of the MASSW dataset, we focus on Computer Science publications from 17 top-tier conferences listed in CSRankings.org, which we identify as relevant to the broader field of AI. We access the publications through Open Academic Graph (OAG)[1], a linked graph database for academic entities including publications, venues, affiliations, and authors (Zhang et al., 2022, 2019a). In total, 191,055 papers that span from 1969 to 2024 are collected, among which, 152,027 contain both a title and an abstract. More details about data curation can be found in Appendix A.

Automatic aspect summarization with LLMs. Most relevant datasets on structured summaries of publications were created using human annotations, which only cover tens to thousands of papers (Mei and Zhai, 2008; Fisas et al., 2015; Cachola et al., 2020b; Wang et al., 2022; Takeshita et al., 2024), limiting their scope for depicting a broader scientific research landscape. For MASSW, we leverage the power of LLMs (e.g., GPT-4) to automatically summarize the five core aspects for all collected papers that have a title and an abstract. In brief, the prompt contains the same content as the annotation guidelines provided to human annotators, along with an one-shot example. More details of LLM-based summarization, including the prompts used, are described in Appendix B.

[1]The OAG dataset is publicly released under the ODC-BY license.

	#Papers with	Avg. #Tokens
Abstract	152,027	145.3
Context	149,849	34.8
Key Idea	149,411	35.1
Method	142,241	30.7
Outcome	132,614	27.6
Projected Impact	72,983	27.2
All Aspects	62,506	N/A

Table 2: Basic statistics of MASSW.

2.3 Dataset Statistics and Visualization

Table 2 reports basic statistics of the MASSW dataset and each of the aspects. We include a visualization of context in Figure 1 in Appendix C to demonstrate the wide landscape captured by MASSW.

3 Dataset Validation

Are LLM-generated summaries trustfully describing the core aspects of the scientific workflow? We validate the structured summaries in MASSW by comparing them with human-generated summaries. We curate a small-scale subset of publications and solicit the human annotations of the same five aspects. This subset demonstrates the alignment between the LLMs and human experts in generating the multi-aspect summary of scientific workflows.

3.1 Evaluation Metric

We employ two categories of similarity evaluation metrics: lexical-based and semantic-based. Lexical-level metrics, such as *BLEU* (Papineni et al., 2002) and *ROUGE*[2] (Lin, 2004), imposing strict requirements on lexical similarity, are prevalent across various natural language generation tasks like machine translation. Nevertheless, studies indicate their limited alignment with human judgments, primarily due to their reliance on exact word matches (Sellam et al., 2020; Callison-Burch et al., 2006; Ananthakrishnan et al., 2006; Sai et al., 2022). Conversely, semantic-based metrics represent a more nuanced perspective, assessing the similarity in meaning or content through the use of pre-trained language models. We utilize four semantic metrics: *BERTScore (BS)* (Zhang

[2]We report ROUGE-1 that evaluates on unigram.

et al., 2019b), which compares token-wise contextual embeddings, *cosine similarity (CS)*, derived from embeddings generated by (Wang et al., 2024), *BLEURT* (Sellam et al., 2020), which is fine-tuned to reflect human judgment, and *FActScore*, which measures the factual consistency of generated content with the reference text (Min et al., 2023). Their details can be found in Appendix E.

3.2 Evaluation Set and Human Annotation

We use a proportionate stratified sampling method on different venues and publication years to select the annotation subset. Specifically, for each venue, we sorted the papers by publication dates and divided them into 7 equal-sized buckets. From each bucket, we randomly sampled one paper, resulting in a balanced sample of 126 papers that accounts for both venue diversity and temporal distribution.

Two trained human experts who are familiar with reading scientific literature are assigned to annotate the aspects of each paper, based on the title and abstract, following a carefully designed codebook. The complete annotation process is detailed in Appendix D. Table 3 (top) illustrates the agreement between human experts by treating one annotation as the reference and the other as the prediction for each paper[3]. In general, there is a high level of agreement across all five aspects, suggesting that the scientific workflow is well-defined and the annotations do not have obvious individual bias.

3.3 Alignment between MASSW and Human Summaries

Three LLMs are investigated to build the MASSW dataset: GPT-3.5 (OpenAI, 2022), GPT-4 (OpenAI, 2023), and Mixtral 8x7B (Jiang et al., 2024). They are instructed using the same information in the codebook for human annotators, and their generated summaries are evaluated against human annotations, shown in Table 3 and Table 8. Ideally, if the LLM perfectly aligns with human experts, the similarity between an LLM annotation and a human annotation should be comparable to that between the annotations of two humans.

Indeed, for semantic-based metrics, we only see a small difference between LLM-human alignment and human-human agreement, and this pattern is consistent for all three models. This indicates that the semantics of the core aspects of the scientific workflows captured by the LLM closely mirror those by human experts. For lexical-level metrics, there is a more notable disparity, especially between GPT-4 and human experts. Our inspection suggests that this discrepancy primarily arises because GPT-4 tends to generate abstractive summaries, often rephrasing/refining the content contained in the original paper whereas human annotators are inclined to directly quote the narratives in the original paper. This extractive approach is inherently more compatible with lexical-level metrics, which favor direct word overlaps.

Hallucination Not all papers describe all five aspects in the titles/abstracts, implying a risk of hallucinations in extraction with LLMs. In addition to accessing extraction accuracy with human annotations, we also conduct a human evaluation of potential hallucinations. Particularly, we ask the human annotators to mark aspects as "mentioned" or "not mentioned" based on whether the aspect is present in the title/abstract. Evaluation results indicate that GPT-4 has the lowest level of hallucination, with an average recall of "not mentioned" across all aspects at 0.641 (Table 9). In comparison, GPT-3.5 has the highest level of hallucination, with an average recall of 0.204.

Given the desirable alignment with humans and the lowest level of hallucination, we select GPT-4 to extract key aspects for MASSW.

In summary, the MASSW dataset, curated using GPT-4, exhibits a high level of accuracy and a relatively low hallucination rate in identifying the key aspects of scientific workflows from publications.

4 Use MASSW to Benchmark AI4Science Tasks

This section demonstrates how the MASSW dataset can serve as a foundational resource for various AI4Science tasks. As a demonstration, we benchmark multiple off-the-shelf LLMs for a handful of example tasks, and we invite the community to explore the greater variety of downstream tasks. We present the task definitions in Section 4.1 and detail the experimental setups and the performance of baseline methods in Section 4.2

[3] To help understand the scale of these metrics, we include a range of examples with varying levels of similarity in Appendix H

	Aspects	CS	BLEURT	BS	BLEU	ROUGE-1	FActScore
Human Agreement	Context	0.935	0.656	0.942	0.594	0.703	0.974
	Key Idea	0.944	0.618	0.938	0.464	0.637	0.992
	Method	0.900	0.559	0.924	0.357	0.540	0.988
	Outcome	0.936	0.671	0.950	0.608	0.737	0.988
	Projected Impact	0.941	0.742	0.955	0.642	0.748	1.000
GPT 4-Human Alignment	Context	0.940	0.607	0.934	0.384	0.604	0.982
	Key Idea	0.944	0.582	0.928	0.375	0.572	0.990
	Method	0.894	0.510	0.908	0.197	0.450	0.987
	Outcome	0.931	0.603	0.933	0.355	0.596	0.990
	Projected Impact	0.916	0.611	0.933	0.282	0.563	1.000

Table 3: Human agreement (top) and GPT 4-human alignment (bottom) for the five extracted aspects of scientific workflow. CS stands for cosine similarity and BS stands for BERTScore. Human agreement is calculated with one annotation randomly selected as the reference and the other (2 annotations per paper) as the prediction. LLM-human alignment and human-human agreement show a high level of similarity, which indicates that LLMs align well with human experts. Please see Appendix H for a range of examples with varying levels of similarity.

4.1 Task Definitions

We demonstrate two types of downstream tasks that use AI to assist and guide scientific workflows, leveraging the structured nature of the MASSW dataset. These tasks are designed to support researchers in navigating the space of scientific ideas and workflows—serving as an inspiration tool rather than a substitute for human reasoning and decision-making.

- **Workflow Prediction**: A scientific workflow has a sequence of steps, for example, "digesting the literature" → "generating research idea" → "validating the idea" → "interpreting the results" → "planning follow-up research". An effective AI system should assist researchers by extrapolating plausible next steps based on prior steps, thereby offering recommendations that inspire exploration rather than dictating fixed solutions. Hence, for each key aspect in MASSW, we can task a model to make predictions based on the aspects prior in the sequence:
  - IDEA GENERATION: Given the *context* of literature, predict the *key idea* of a new study.
  - METHOD RECOMMENDATION: Given the *context* and a *key idea*, suggest a *method* to validate the idea.
  - OUTCOME PREDICTION: Given the *context*, a *key idea*, and a *method* of validation, forecast the *outcome* of the validation/analysis.
  - FUTURE WORK RECOMMENDATION: Given all other aspects of a study, estimate its *projected impact* and recommend tasks for follow-up studies.
- **Title Prediction**: A subsequent step of the research workflow is to publish the results. A powerful AI copilot should be able to enhance writing by recommending appropriate and appealing titles that encapsulate the key elements of a paper. We therefore introduce the task of title prediction, which challenges an AI model to generate a title given all five aspects of a study.

4.2 Demonstration with Baselines

We now detail the experimental settings and the performance of our baseline models.

- **Test Data**: To create the test set, we employ proportionate stratified sampling based on dates of publication; we select 60 papers (with all aspects mentioned) from each venue to ensure broad representation, resulting in a test set of 1020 papers.
- **Baseline Models**: We test GPT-3.5, GPT-4, and Mixtral 8x7B as baseline models.
- **Prompting Methods**: We test four prompting strategies: (1) zero-shot, (2) zero-shot chain of thought (adding the instruction "*Let's think step by step*" to the end of the zero-shot prompt) (Kojima et al., 2022), (3) few-shot, and (4) few-shot chain of thought (Wei et al., 2022). The models were provided with (i) definitions of all five aspects as defined in Table 1, (ii) all necessary aspects for each task,

and (iii) a specific task instruction. Detailed prompting templates and settings are specified in Appendix I.

- **Evaluation Metrics**: We evaluate the model outputs using the same metrics as described in Section 3. Due to space limit, *BLEURT* and *FActScore* are reported in Table 4, while other metrics are reported in Appendix J.

The benchmarking results are presented in Table 4, offer several interesting observations:

- **Task Complexity:** Among the workflow prediction tasks, outcome prediction and future work recommendation are the most achievable ones with the tested models, although they still present challenges. Outcome prediction often shows higher performance, likely because published work more often reports positive results, making it relatively predictable. Interestingly, future work prediction tends to show a better performance than key idea prediction, even though these two tasks are more homogeneous in nature: both extrapolating new directions from the status quo. This is likely because many papers include only a cursory discussion of "projected" future directions, which tends to be more straightforward and predictable than "real" follow-up research that would lead to a future publication. Idea generation and method recommendation are inherently more difficult, since they require both highly specialized knowledge in the domain and strong innovation capability.

- **Model Performance:** The few-shot prompting method enhances model performance over other methods by helping the models understand the narrative structure and focus required for the tasks. In contrast, adding CoT to zero-shot prompts or using few-shot CoT does not yield significant improvements, indicating that the complexity of the scientific innovation tasks might exceed the reasoning capabilities of the off-the-shelf LLMs without further instructing and fine-tuning in the particular scientific domain.

- **Influence of Metrics:** Evaluation metrics also play critical roles in producing the results. We observe that GPT-4 consistently outperforms GPT-3.5 and Mixtral-8x7B when evaluated by BLEURT. However, Mixtral-8x7B excels when evaluated by FActScore. This reveals the complexity of evaluating scientific workflows, as different metrics can measure different aspects of generated content. Further research is needed to understand these nuances.

Overall, our experiments offer a demonstration of several downstream AI4Science tasks that can be facilitated by the MASSW dataset, and they highlight the complexities and nuances of integrating AI models into scientific workflows. The demonstrated tasks are by no means the complete set, and the benchmarked models are by no means the best ones. With our dataset, additional tasks of AI-assisted scientific discovery can be designed, and additional AI/machine learning models can be tested and optimized. Readers may consider using part of this dataset to test various instructing/prompting strategies, fine-tune LLMs for scientific reasoning, or implement retrieval-augmented solutions. Furthermore, the current evaluation metrics, which primarily assess semantic and lexical similarity, may not adequately reflect the nuances of the tasks. AI models could have generated meaningful ideas or research methods completely different from what's reported in the original paper. More sophisticated evaluation procedures or metrics could be advantageous given the rich and structured information in MASSW.

5 Related Work

Aspect-based document summarization Aspect-based document summarization generates summaries focused on specific document aspects rather than providing a general overview. These aspects may be predefined (E et al., 2023; Santosh et al., 2024; Frermann and Klementiev, 2019; Takeshita et al., 2024; Fisas et al., 2015) or dynamically determined based on content (Amar et al., 2023; Xu et al., 2011; Coavoux et al., 2019; Yang et al., 2023; Hayashi et al., 2021). In our case, the aspects are predefined with domain knowledge, identifying five major aspects inherent in scientific workflows. Aspect-based summarization has been widely applied across various domains. For instance, in the legal domain, (Santosh et al., 2024) developed a challenging dataset for summarizing legal case decisions. In the context of online shopping, (Coavoux et al., 2019) and (Xu et al., 2011) have explored the

Model	Prompt	BLEURT Score					FActScore				
		Idea	Method	Outcome	Future	Title	Idea	Method	Outcome	Future	Title
GPT-3.5	0-Shot	0.413	0.384	0.406	0.411	0.455	0.741	0.873	0.895	0.931	0.819
	2-Shot	0.411	0.389	0.421	0.443	**0.471**	0.833	0.877	0.931	0.932	**0.936**
	0-CoT	0.340	0.367	0.395	0.422	0.442	0.823	0.832	0.836	0.893	0.713
	2-CoT	0.396	0.382	0.399	0.443	0.447	0.846	0.869	0.908	0.933	0.808
GPT-4	0-Shot	**0.435**	0.390	0.420	**0.456**	0.442	0.744	0.805	0.822	0.893	0.628
	2-Shot	0.421	**0.400**	**0.440**	0.431	0.460	0.725	0.777	0.837	0.905	0.786
	0-CoT	0.412	0.395	0.410	0.451	0.441	0.812	0.828	0.853	0.907	0.696
	2-CoT	0.412	0.373	0.431	0.421	0.439	0.809	0.832	0.875	0.941	0.725
Mixtral-8x7B	0-Shot	0.329	0.328	0.340	0.367	0.343	0.873	0.940	0.942	0.971	0.788
	2-Shot	0.326	0.312	0.327	0.369	0.385	0.894	**0.952**	**0.950**	0.977	0.850
	0-CoT	0.297	0.327	0.317	0.351	0.343	0.796	0.917	0.854	0.954	0.862
	2-CoT	0.386	0.349	0.383	0.417	0.396	**0.932**	0.944	0.944	**0.985**	0.852
Average		0.382	0.366	0.391	0.415	0.426	0.819	0.871	0.887	0.935	0.789

Table 4: Evaluation results of the five benchmark tasks: Idea Generation ("Idea"), Method Recommendation ("Method"), Outcome Prediction ("Outcome"), Future Work Recommendation ("Future"), and Title Prediction ("Title"). k-CoT stands for k-shot CoT. The models with the best performance are **bolded**.

.

dynamic generation of multiple aspect-based summaries for online reviews. Related to our work, (Mei and Zhai, 2008), (Takeshita et al., 2024), and (Fisas et al., 2015) have created annotated datasets for summarizing publications in information retrieval, natural language processing, and computer graphics, respectively.

Our approach differs significantly in scope and objective from such studies. In these researches, the primary goal is to establish benchmarks for evaluating models' summarization capabilities, and therefore their end product is usually a limited set of human-annotated examples. In contrast, we aimed to develop a comprehensive, large-scale dataset of scientific workflows, where LLMs, after validation, are used as a proxy for human experts to generate the dataset. Our purpose in creating this dataset is to support extensive downstream tasks that employ AI to assist scientific innovation (such as key idea generation). This purpose is achieved by including a much larger volume of scientific publications and tailoring the definition of aspects so that they are closely tied to the exploration of scientific workflows.

Scientific workflow automation With the rise of LLMs and autonomous agents, many studies investigate the potential of using LLM agents to engage with certain components of scientific workflows, traditionally managed solely by human researchers. Huang et al. (2024) proposes using domain-knowledge-augmented LLM agents to automate and enhance the design of CRISPR-based gene-editing experiments. Liu et al. (2024) finds that GPT-4 is useful in converting experimental workflow ideations into executable code on microscope APIs. Boiko et al. (2023) shows that an AI system driven by GPT-4 can autonomously design, plan, and execute multiple complex experiments in chemical syntheses. Agarwal et al. (2024) provides an LLM-based toolkit for reviewing scientific literature on a given topic, utilizing retrieval augmented generation to access the latest research. Procko et al. (2023) uses LLMs to enhance scientific writing by creating a taxonomy of paper structures, thereby improving efficiency in the academic publishing pipeline. These existing studies focus on particular scientific domains and specific use cases, while a significant gap remains in systematically measuring the effectiveness of LLM agents in planning and navigating scientific workflows in general. Our work addresses this gap by introducing multiple new benchmark tasks that assess the capabilities of LLMs across various essential stages of the scientific reserch process.

6 Conclusion

We present MASSW, a comprehensive dataset that structures and summarizes the extensive under-organized scientific literature in computer science. By leveraging LLMs to extract five core aspects from over 152,000 publications, we have aligned these summaries with critical steps in the scientific workflow. Our validation confirms the high quality of these summaries, demonstrating their utility in facilitating and benchmarking various downstream tasks and analyses. MASSW serves as a valuable and trustworthy resource for developing new AI methods that optimize scientific workflows and foster innovation. We anticipate that this dataset will enable researchers to more effectively navigate scientific literature and inspire future advancements in AI-driven scientific discovery.

7 Limitations

Offline Evaluation of Scientific Workflow Prediction Scientific workflow prediction, like many other text generation tasks, is inherently open-ended. A given context can lead to multiple plausible outcomes, such as several valid research ideas emerging from the same information. Rigorous evaluation of these outcomes typically requires rich feedback from domain experts (Si et al., 2024), which limits the scalability and thus prevents automating the process as a large-scale benchmark. While our current evaluation setup focuses on comparing the predictions with the ground truth by measuring the textual similarity, it offers an offline, large-scale, and automated benchmarking solution that is able to cover a wide range of scientific fields. Future work could explore incorporating more comprehensive evaluation techniques to enable a more nuanced assessment of scientific workflow prediction.

Limitations of Data Sources Another limitation of MASSW is that it derives scientific workflows solely from the titles and abstracts of publications, primarily due to limited access to full paper texts and the costs associated with processing large volumes of full papers. While incorporating full texts—including tables and figures—could enhance the accuracy and comprehensiveness of workflow summaries, this approach faces several challenges. Copyright restrictions limit access to a substantial portion of full texts, and extracting structured information from PDF-only formats presents technical difficulties that may compromise data quality. Additionally, the processing costs would be significantly higher, with our initial GPT-4 API expenses already reaching approximately $7,500. Given these constraints, we acknowledge this limitation and have listed it as a future direction. In an updated version of MASSW, we plan to release full-text summaries for a subset of papers. Expanding the dataset to include literature from other domains beyond AI-related computer science conferences is another potential avenue for improvement, ensuring broader applicability of the defined aspects across different scientific fields.

Potential Societal Impacts We anticipate that MASSW would unleash the great potential of building AI tools to optimize scientific workflows and therefore accelerate the progress in AI for Science. We are, however, aware of two potentially negative impacts. First, by the selection of top-tier AI-related conferences, MASSW might introduce biases to the downstream AI copilots, potentially diminishing the influence of other venues and limiting the diversity of research topics and methodologies considered. This may be addressed by iterating and expanding the scope of MASSW. Second, the reliance on AI-generated summaries and recommendations could lead researchers to depend on these tools, reducing their engagement with the original papers, and therefore overlook the nuances documented in the literature. This remains an open question for human-AI collaboration.

Broader Applications Beyond Machine Learning While the demonstrated downstream tasks primarily focus on machine learning prediction, MASSW offers many other valuable applications. First, MASSW can serve as a resource for measuring the novelty of scientific ideas within the context of historical knowledge (Wang et al., 2023). Second, it can aid in building search engines that allow researchers to navigate fine-grained scientific workflows. For example, one could retrieve ideas based on similar problem contexts or methods used in related studies, facilitating the ideation process. Third, MASSW enables the visualization of scientific literature and workflows, supporting large-scale bibliometric analyses in the scientific domain.

References

Shubham Agarwal, Issam H Laradji, Laurent Charlin, and Christopher Pal. 2024. LitLLM: A toolkit for scientific literature review.

Shmuel Amar, Liat Schiff, Ori Ernst, Asi Shefer, Ori Shapira, and Ido Dagan. 2023. OpenAsp: A benchmark for multi-document open aspect-based summarization.

R Ananthakrishnan, P Bhattacharyya, M Sasikumar, and Ritesh M Shah. 2006. Some issues in automatic evaluation of English-Hindi MT: More blues for BLEU. *Icon*.

Francisco J Ayala. 2009. Darwin and the scientific method. *Proceedings of the National Academy of Sciences*, 106(supplement_1):10033–10039.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations : ambiguity, variation, uncertainty, error and bias. In *14th Linguistic Annotation Workshop (LAW 14)*, pages 60–73. Association for Computational Linguistics.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020a. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020b. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised Aspect-Based Multi-Document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. *arXiv preprint arXiv:1909.04054*.

Franck Dernoncourt and Ji Young Lee. 2017. Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. *arXiv preprint arXiv:1710.06071*.

Nikhil E, Anshul Padhi, Pulkit Parikh, Swati Kanwal, Kamalakar Karlapalem, and Natraj Raman. 2023. Aspect-based summarization of legal case files using sentence classification. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, pages 1249–1252, New York, NY, USA. Association for Computing Machinery.

Beatriz Fisas, Horacio Saggion, and Francesco Ronzano. 2015. On the discoursive structure of computer graphics research papers. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 42–51, Denver, Colorado, USA. Association for Computational Linguistics.

Raymond Fok, Hita Kambhamettu, Luca Soldaini, Jonathan Bragg, Kyle Lo, Marti Hearst, Andrew Head, and Daniel S Weld. 2023. Scim: Intelligent skimming support for scientific papers. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 476–490.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A dataset for multi-domain aspect-based summarization. *Trans. Assoc. Comput. Linguist.*, 9:211–225.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2018. Accelerating innovation through analogy mining. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, California. International Joint Conferences on Artificial Intelligence Organization.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. 2024. CRISPR-GPT: An LLM agent for automated design of Gene-Editing experiments.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yongtao Liu, Marti Checa, and Rama K Vasudevan. 2024. Synergizing human expertise and AI efficiency with language model for microscopy operation and automated experiment design.

Qiaozhu Mei and ChengXiang Zhai. 2008. Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pages 816–824.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E Miller, Maddie Simens, Amanda Askell, P Welinder, P Christiano, J Leike, and Ryan J Lowe. 2022. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.*, abs/2203.02155.

K Papineni, Salim Roukos, T Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. *Annu Meet Assoc Comput Linguistics*, pages 311–318.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Jackson Sargent, Apostolos Dedeloudis, and David Jurgens. 2022. POTATO: The portable text annotation tool.

Tyler Procko, Alexandra Davidoff, Timothy Elvira, and Omar Ochoa. 2023. Towards improved scientific knowledge proliferation: Leveraging large language models on the traditional scientific writing workflow. *SSRN Electron. J.*

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):1–39.

T Y S Santosh, Mahmoud Aly, and Matthias Grabmair. 2024. LexAbSumm: Aspect-based summarization of legal decisions.

T Sellam, D Das, and A P Parikh. 2020. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Paolo Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications.

Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. 2016. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297.

Eberhard O Voit. 2019. Perspective: Dimensions of the scientific method. *PLOS Computational Biology*, 15(9):e1007279.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Xueke Xu, Tao Meng, and Xueqi Cheng. 2011. Aspect-based extractive summarization of online reviews. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, SAC '11, pages 968–975, New York, NY, USA. Association for Computing Machinery.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-Scale open domain aspect-based summarization. pages 4381–4401.

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Evgeny Kharlamov, Bin Shao, et al. 2022. Oag: Linking entities across large-scale heterogeneous knowledge graphs.

IEEE Transactions on Knowledge and Data Engineering.

Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019a. Oag: Toward linking large-scale heterogeneous entity graphs. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2585–2595.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. BERTScore: Evaluating text generation with BERT.

A Data Curation

To build this initial version of the MASSW dataset, we focus on Computer Science publications from 17 top-tier conferences listed in CSRankings.org, which we identify as relevant to the broader field of AI. Below we list the conferences included in MASSW.

- Artificial Intelligence: AAAI, IJCAI;
- Computer Vision: CVPR, ECCV, ICCV;
- Machine Learning: ICLR, ICML, NeurIPS, KDD;
- Natural Language Processing: ACL, EMNLP, NAACL;
- The Web & Information Retrieval: SIGIR, WWW;
- Databases: SIGMOD, VLDB;
- Interdisciplinary Areas: CHI.

We access the publications through Open Academic Graph (OAG)[4], a linked graph database for academic entities including publications, venues, affiliations, and authors (Zhang et al., 2022, 2019a). For publications before the year 2020, we access the data through OAG v2.1[5], which is generated in 2020 and contains publications as early as 1969. For publications in and after 2020, we access the data through OAG v3.1[6], which is generated in Feb, 2024 and contains publications from 2000 to 2024.

In total, 191,055 papers are collected that span from 1969 to 2024, among which, 152,027 contain both a title and an abstract.

B Aspect Summarization

We use OpenAI GPT-4 (snapshot `gpt-4-0613`[7]) to summarize the core aspects of each collected publication. Here we provide the prompt we used for automated summarization:

[4]The OAG dataset is publicly released under the ODC-BY license.

[5]OAG v2.1: https://old.aminer.cn/oag-2-1/oag-2-1.

[6]OAG v3.1: https://open.aminer.cn/open/article?id=65bf053091c938e5025a31e2.

[7]OpenAI GPT-4 models: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4.

Aspect Summarization Prompt

System message:
Instructions
You are an expert in computer science. Your task is to summarize the following five aspects of the papers given the definitions below.

Definitions of Aspects
Context
- The status quo of related literature or reality which motivated this study. This could normally be a problem, a research question, or a research gap that has not been successfully addressed by previous work.
- Anything happened before this study.
Key Idea
- The main intellectual merit of this paper, often in comparison to the context. This could normally be a novel idea or solution proposed in this paper that distinguishes it from what's already done in literature.
- Proposed in this study.
Method (Validation Methodology)
- The specific experiment or proof that investigates and validates the key idea.
- CS papers often refer "Method" as algorithm or model, but our definition here is **different**.
- Performed in this study.
Outcome
- The factual statement about the study output. This could be the experiment results and any other measurable outcome that has occurred. It marks whether the key hypothesis is testified or not.
- Produced in this study.
Future Impact
- The impact of the work on the field explicitly anticipated by the authors, and potential further research explicitly identified by the author that may improve or extend this study.

Notes
- If an aspect is NOT mentioned in the abstract, mark it as "N/A" (not applicable). DO NOT come up with your own interpretation.
- Each aspect should be summarized in 1-2 sentences in most cases.
- Each aspect should be self-contained and should not contain references including other aspects (cross-reference).
- Including specific names of proposed models, datasets, etc., in the summary is acceptable.
- If the problem definition is novel (e.g., proposing a new task), classify it as a Key Idea.
- Non-measurable outcomes should be categorized as Future Impact.
- Impacts that have already occurred should be considered as Outcome.
- A new observation that motivates the proposal of a key idea should be classified under Key Idea.
- Future Impact should not account for real impacts, such as the number of citations a paper has received.

Aspect Summarization Prompt (Cont'd)

Response Format
The response should be a JSON object in the following format:
```json
{
"Context": "...",
"Key Idea": "...",
"Method": "...",
"Outcome": "...",
"Future Impact": "..."
}
```

User message:
Title: Attention Is All You Need
Abstract: The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Assistant message:
{
"Context": "The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing sequence transduction models connect the encoder and decoder through an attention mechanism.",
"Key Idea": "The authors propose a simple network architecture called Transformer based solely on attention mechanisms and dispenses with recurrence and convolutions.",
"Method": "The authors perform experiments on the WMT 2014 English-to-German and English-to-French translation task. The authors apply the proposed model to English constituency parsing both with large and limited training data.",
"Outcome": "The proposed model achieves a BLEU score of 28.4 on the WMT 2014 English-to-French translation task. The proposed model achieves a BLEU score of 41.8 on the WMT 2014 English-to-German translation task after training for 3.5 days on 8 GPUs.",
"Future Impact": "N/A"
}

User prompt:
Title: [The title of the publication to be summarized]
Abstract: [The abstract of the publication to be summarized]

In the prompt, the first user message and the assistant message provide a one-shot example to the LLM. During automated summarization, GPT-4 failed to summarize 12 publications due to safety filtering. In total, we obtained core aspects from 152,015 publications.

C Dataset Visualization

Figure 1 visualizes the *context* aspect as an example. We embed the aspect summary using OpenAI Ada model, then conduct dimension reduction with LargeVis (Tang et al., 2016). We then use BERTopic to assign labels to each cluster identified by HDBSCAN.

D Human Annotation Process

D.1 Overview

To ensure high-quality annotations, we recruited five student researchers from the University of Michigan, all with verified backgrounds in AI. Each paper was independently annotated by two different researchers to minimize individual bias and ensure annotation reliability. We conducted the annotation process using the Potato annotation platform (Pei et al., 2022), providing annotators with only the paper titles and abstracts to maintain consistency in the input data.

D.2 Codebook

Task Description Our task is to construct a dataset for multi-aspect summarization of scientific papers. Our papers of interest are from top computer science conferences. For each paper, the aspects of interest include the following: (Same content in Table 1)

Your task is to write summarizations of these five aspects for each paper assigned to you. We have the following requirements for this task:

- Read the content thoroughly before writing your summaries.
- Write a short summary for each aspect (1-2 sentences in most cases).
- Each aspect should be self-contained and should not contain references including other aspects (cross reference).
- Only consider the abstract section and title as the input.

FAQ

- **Q:** Is it fine to include the specific name of the proposed model/dataset/etc in the summary?
  A: Yes, it is fine to include them.
- **Q:** If the problem definition is novel (i.e. proposing a new task), should it be a key idea or context?
  A: Key idea.
- **Q:** If the concept is not mentioned at all in the abstract, what should I do?
  A: Mark it as "N/A" (not applicable).
- **Q:** If the author claims a non-measurable outcome, should it be considered as an Outcome or Future Impact?
  A: Future Impact.
- **Q:** If the author mentions an impact that has happened (e.g. the first work to ...), should it be considered as an Outcome or Future Impact?
  A: Outcome.
- **Q:** If the author mentions a new observation that motivates them to propose the key idea, should it be considered as context or key idea?
  A: Key idea.
- **Q:** Should future impact consider its real impact? For example, a paper gains a lot of citations.
  A: Future Impact should not consider other papers.

E Implementation Details of Semantic-Based Evaluation Metrics

- *Cosine Similarity*: We compute the cosine similarity between sentence embeddings generated by `multilingual-e5-large-instruct` from HuggingFace.
- *BLEURT*: We use the pre-trained checkpoint `BLEURT-20-D12`.
- *BERTScore*: We use the pre-trained checkpoint from HuggingFace https://huggingface.co/spaces/evaluate-metric/bertscore.

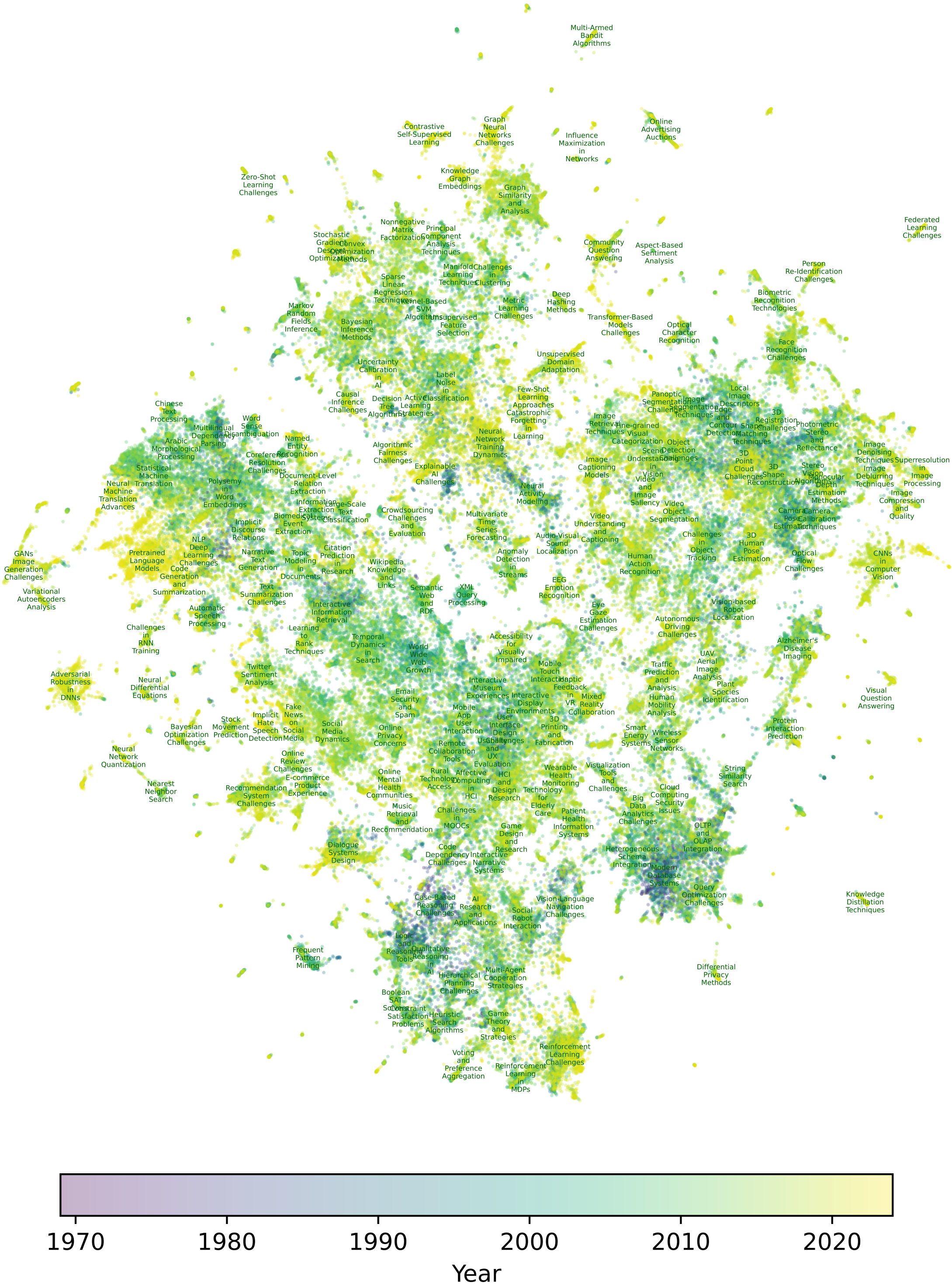

Figure 1: Low dimensional visualization of contexts.

- *FActScore*: We use the official FActScore implementation with some modifications. Specifically, while the original FActScore retrieves references from a trusted corpus, we adapt the approach by directly providing the reference following `https://github.com/shmsw25/FActScore/issues/32`.

We follow the implementation of ROUGE to select the maximum score when there are multiple references.

F Annotation Set Size Validation

We also conduct analysis to guarantee the sample size of 126 papers is sufficient to evaluate the quality of the model-extracted aspects. We report the results of statistical tests comparing human-human agreement and model-human agreement using t-tests in Table 6. The agreements measured by semantic similarity (CosineSimilarity, BLEURT, and BERTScore) are mostly not significantly different under the t-test (generally, $p > 0.05$). The agreements measured by lexical similarity (BLEU and ROUGE-1) are statistically different, which is expected as humans and the LLM tend to use different words.

G Tables for LLM-human agreement and hallucination experiments.

In Table 8, we show the results of LLM-human agreement for GPT-3.5 and Mixtral-8x7B. Both models demonstrate a high level of alignment with human annotation. Table 9 includes the results of hallucination experiments. Each cell presents the recall of "not mentioned" aspects as identified by human experts. Higher values indicate lower rates of hallucination.

Table 7 shows the breakdown of the "not mentioned" ratio identified by human annotators and the false-positive ratio of GPT-4 (where an aspect is "not mentioned" but falsely summarized by the LLM due to hallucination). All but one aspects are missing in no more than 20% of the data according to human annotation. Every abstract mentions the key idea, and over 90% mention outcomes. The only aspect not mentioned in a significant portion of abstracts is "projected impact," as described in Table 2. Notably, GPT-4 achieves strong recall (0.923) in identifying missing "projected impact" cases. As a result, the actual false-positive ratio of GPT-4 remains below 10% for all five aspects.

H Examples of Texts for Different Similarity Levels

We provide two examples of texts to illustrate how the evaluation metrics could be interpreted. The evaluation results can be found at Table 10 and 11.

- **Reference 1:** InstructGPT, even with 100x fewer parameters, is preferred over GPT-3 in human evaluations. It shows improvements in truthfulness and reductions in toxic outputs with minimal performance regressions on public NLP datasets.
- **Example 1a:** InstructGPT, despite having 100x fewer parameters, is preferred over the larger GPT-3 according to human evaluations, demonstrating better truthfulness and fewer toxic outputs with only minimal regressions in performance on public NLP benchmarks.
- **Example 1b:** Human evaluations favor the 1.3B parameter InstructGPT model over the 175B GPT-3 model, even though it has significantly fewer parameters. It also shows enhanced truthfulness and reduced generation of toxic content, with negligible declines in performance across standard NLP datasets.
- **Example 1c:** In human assessments, the smaller InstructGPT model, which has far fewer parameters, outperforms GPT-3, showing not only increased accuracy but also less toxic output, with only slight performance downturns on widely recognized NLP tests.
- **Example 1d:** This paper explores the enhancement of language model alignment with human intent through fine-tuning methods using labeler feedback and reinforcement learning, resulting in a smaller, more efficient model that surpasses a much larger baseline in both user satisfaction and safety metrics.
- **Example 1e:** Effective communication is not about speaking more; it's about achieving more with fewer words.

- **Reference 2:** The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an

	Context	Key Idea	Method	Outcome
CS	(-1.346, 0.180)	(-0.069, 0.945)	(-0.911, 0.364)	(0.148, 0.882)
BLEURT	(1.711, 0.089)	(1.845, 0.066)	(1.388, 0.167)	(2.736, 0.007)
BERTScore	(1.427, 0.155)	(2.140, 0.033)	(1.191, 0.235)	(3.077, 0.002)
BLEU	(4.548, 0.000)	(2.968, 0.003)	(2.899, 0.004)	(7.357, 0.000)
ROUGE-1	(2.936, 0.004)	(2.418, 0.016)	(2.103, 0.037)	(4.268, 0.000)

Table 6: Statistical tests comparing human-human agreement and model-human agreement using t-tests. Numbers in parentheses are t values followed by p values.

Aspect	Human NA (%)	GPT-4 FP (%)
Context	20.0	8.3
Key Idea	0.0	0.0
Method	15.8	9.1
Outcome	9.2	3.3
Projected Impact	86.7	6.7

Table 7: Analysis of aspects not mentioned in abstracts and GPT-4's false-positive rates.

encoder-decoder configuration. The best performing sequence transduction models connect the encoder and decoder through an attention mechanism.

- **Example 2a:** The leading sequence transduction models utilize complex recurrent or convolutional neural networks in an encoder-decoder framework, with the most effective models incorporating an attention mechanism between the encoder and decoder.

- **Example 2b:** Traditional sequence transduction models rely on sophisticated recurrent or convolutional neural networks arranged in an encoder-decoder setup, where top-performing models are distinguished by the use of an attention mechanism linking the encoder and decoder.

- **Example 2c:** Existing high-performing sequence transduction models typically feature either recurrent or convolutional neural networks configured in an encoder-decoder structure, often enhanced with an attention mechanism to improve performance.

- **Example 2d:** The paper introduces the Transformer, a novel network architecture that eschews recurrent and convolutional structures in favor of a design entirely based on attention mechanisms, aiming to enhance parallelizability and reduce training time.

- **Example 2e:** "Simplicity is the ultimate sophistication." - Leonardo da Vinci

I Experiment Details

Test set sampling. In the benchmark section, We use proportionate stratified sampling to construct the test set. According to publication year, we separate the year range into at most 10 strata (i.e. groups). Each group covers approximately the same number of years. The we sample from each strata proportionally to the number of papers in that strata. The number of samples for each venue is 60, which results in 1020 papers in total.

Prompting templates. We recall that the model will take in three part of information: (i) definitions of all five aspects, (ii) all necessary aspects for each task, and (iii) a specific task instruction. We include the prompts for all tasks below.

	Aspects	CS	BLEURT	BS	BLEU	ROUGE-1	FactScore
Human Agreement	Context	0.935	0.656	0.942	0.594	0.703	0.974
	Key Idea	0.944	0.618	0.938	0.464	0.637	0.992
	Method	0.900	0.559	0.924	0.357	0.540	0.988
	Outcome	0.936	0.671	0.950	0.608	0.737	0.988
	Projected Impact	0.941	0.742	0.955	0.642	0.748	1.000
GPT-3.5-Human Alignment	Context	0.934	0.597	0.934	0.524	0.635	0.989
	Key Idea	0.936	0.575	0.927	0.439	0.582	0.980
	Method	0.895	0.510	0.910	0.197	0.445	0.984
	Outcome	0.928	0.608	0.934	0.452	0.626	0.986
	Projected Impact	0.876	0.498	0.905	0.170	0.371	0.989
Mixtral-8x7B-Human Alignment	Context	0.944	0.645	0.946	0.590	0.693	0.971
	Key Idea	0.949	0.636	0.943	0.556	0.662	0.973
	Method	0.905	0.554	0.920	0.295	0.509	0.987
	Outcome	0.933	0.674	0.948	0.665	0.707	0.991
	Projected Impact	0.917	0.635	0.936	0.384	0.599	0.987

Table 8: GPT-3.5 and Mixtral-8x7B-human alignment.

Model	Context	Method	Outcome	Projected Impact
GPT-3.5	0.000	0.105	0.364	0.346
GPT-4	**0.583**	**0.421**	**0.636**	**0.923**
Mixtral-8x7B	0.042	**0.421**	0.364	0.750

Table 9: Recall of "not mentioned" aspects as identified by human experts. Higher values indicate lower rates of hallucination. Key idea is not included as it presents in all papers in the annotation set.

Example	CS	BLEURT	BS	BLEU	ROUGE-1
1a	0.9500	0.7185	0.9589	0.2753	0.6970
1b	0.9366	0.6202	0.9188	0.0000	0.5135
1c	0.9326	0.5572	0.9109	0.0000	0.3582
1d	0.8384	0.3119	0.8504	0.0000	0.1351
1e	0.7594	0.1953	0.8396	0.0000	0.1702

Table 10: Evaluation of similarity between examples and Reference 1 using various metrics.

Example	CS	BLEURT	BS	BLEU	ROUGE-1
2a	0.9572	0.7256	0.9613	0.3772	0.7077
2b	0.9516	0.6781	0.9494	0.2689	0.6857
2c	0.9381	0.5660	0.9289	0.1927	0.5079
2d	0.8355	0.3598	0.8645	0.0000	0.2687
2e	0.7108	0.1728	0.8095	0.0000	0.0476

Table 11: Evaluation of similarity between examples and Reference 2 using various metrics.

Prompt

System message:
You are an expert in research tasked with generating detailed prompts for various aspects of academic research papers. Each task involves creating a specific type of prompt based on the provided information. Here are the definitions of each part you will work with:
- Concept
    - Definition
    - Relative Time
- <Definitions of Context, Key Idea, ...>

Prompt

Template for idea generation:
Given the context: '{context}', generate key ideas that could advance this area of study.

Template for method recommendation:
Given the context: '{context}' and the key idea: '{key_idea}', recommend the most suitable method to validate this idea.

Template for outcome prediction:
Based on the context: '{context}', the key idea: '{key_idea}', and the recommended method: '{method}', predict the potential outcome of this research.

Template for impact prediction:
Based on the context: '{context}', the key idea: '{key_idea}', the method: '{method}', and the outcome: '{outcome}', suggest projected Impact for this research.

Template for title prediction:
Given the context: '{context}', the key idea: '{key_idea}', the method: '{method}', the outcome: '{outcome}', and the future impact: '{future_impact}', predict the title of this research paper. The title should be concise and reflective of the core aspects.

Details about prompting methods. For the zero shot prediction, model will take in the system prompt and user prompt. For the few shot prompting, we add two fixed round of conversation before the actual user request. The few-shot examples can be found in the code-base under the data folder. For chain of thought prompts, we add the sentence "Let's think step by step. The final prediction should start after the marker 'Prediction:'." at the end of zero-shot prompts. After LLMs produce the output, we extract the content after the word "Prediction" as the final prediction.

Details about baseline LLMs. We use Azure platform for all baseline models. The model id used in this study are `gpt-35-turbo` `(0125)`[8], `gpt-4 (turbo-2024-04-09)` and `mistralai-mixtral-8x7b-instru-7`[9]. For the mistral model, as the model does not take in system prompt, we replace it with one round of conversation (user: <system prompt>, assistant: `I got it. Please give me further instructions!`).

Example model outputs. We include an example output for all the prompting methods in Table 12.

J Supplementary Tables

We include the benchmark results for Cosine Similarity (CS), BERTScore (BS), BLEU and ROUGE-1 in Table 13, 14, 15, 16 respectively.

K Downstream Tasks Evaluation with Human Annotated Data

To demonstrate if the LLM-generated summarization influence the evaluation of LLMs under the downstream tasks, we further evaluate the models with human annotated summaries. We report the result in Table 17-18. Overall, the observed trends remain consistent with respect to task complexity, model performance, and the influence of metrics, demonstrating the robustness of our findings.

L Dataset Documentation and Intended Uses

We use the Data Cards recommended by the submission guideline. Please refer to our anonymous code repo.

M Author Statement

All the authors bear all responsibility in case of violation of rights, etc., and we confirm the data license.

[8] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models

[9] https://learn.microsoft.com/en-us/azure/machine-learning/how-to-deploy-models-mistral?view=azureml-api-2&tabs=mistral-large

Prompt Type	Model Output/Ground Truth
Ground Truth	This paper introduces the concept of co-space applications, which provide opportunities unavailable in separate physical or virtual spaces, and posit that the database community can significantly contribute to this emerging field.
0-shot	**Key Idea 1: Development of a Unified Interaction Framework** - **Concept:** This key idea proposes creating a unified interaction framework that enables users to interact with both physical and virtual elements through a single interface or set of interaction gestures. This framework integrates sensory feedback mechanisms (e.g., haptic, auditory, visual cues) that are context-aware and adapt based on the user's actions and environmental conditions. - **Relative Time:** Proposed in this study. **Key Idea 2: Hybrid Space Adaptation Engine** - **Concept:** The introduction of a hybrid space adaptation engine designed to dynamically modify the virtual space in response to changes in the physical environment and vice versa. This engine uses real-time data analytics and machine learning algorithms to predict and adapt to user needs and environmental shifts, ensuring a consistent and immersive experience across both domains. - **Relative Time:** Proposed in this study. <Some other key ideas.>
2-shot	The researchers propose the development of an intuitive Mixed Reality (MR) interface that leverages advancements in augmented reality (AR) and virtual reality (VR), combined with real-time spatial computing and machine learning algorithms. This interface is designed to enable more seamless and intuitive user control and interaction between physical and virtual spaces, enhancing the co-space experience by making digital manipulations feel as tangible as physical ones.
CoT	To address the context effectively, a detailed step-by-step approach will be taken to generate key ideas that could significantly advance the study of co-space environments, where the integration and interaction between physical and virtual spaces present both challenges and opportunities for technological innovation. These key ideas are crafted to facilitate a seamless manipulation of real and digital spaces by users. ### Concept: - **Definition**: Co-space environments are settings in which physical and digital realms are intertwined, allowing for continuous interaction and manipulation across these domains. - **Relative Time**: This concept is becoming increasingly relevant with the advances in augmented reality (AR), virtual reality (VR), mixed reality (MR), and Internet of Things (IoT) technologies. ### Key Ideas: 1. **Development of a Unified Interaction Protocol (UIP)** - This protocol would standardize the way in which interactions occur between the physical and virtual components of a co-space environment. The UIP would ensure that devices, regardless of their operational technology (OT) or information technology (IT) origins, can communicate and interact seamlessly in a common language. 2. **Real-time Spatial Mapping and Augmentation Technology** - An advanced spatial mapping technology that dynamically captures the physical environment and accurately overlays digital information in real-time. This would not only facilitate seamless user manipulation of digital objects in real spaces but also enhance the realism and interactiveness of the virtual space. <Some other key ideas.> ### Prediction: - The aforementioned key ideas, if successfully implemented, are poised to revolutionize the co-space environment by breaking down the barriers between the physical and virtual realms. These innovations would not only facilitate seamless manipulation of spaces by users but would also set a new standard for interactive technologies, significantly enhancing user experiences while addressing privacy and security concerns. Moreover, the advancement in such areas could pave the way for novel applications ranging from immersive education and remote collaboration to advanced urban planning and beyond.
2-shot CoT	Integrating physical and virtual realms in co-space environments poses technical and usability challenges, necessitating novel solutions for seamless interaction. One way forward is to bridge the gap between these spaces with technologies that can intuitively understand and manipulate both contexts in real-time. Prediction: The authors propose an innovative interaction framework that blends Augmented Reality (AR) and Internet of Things (IoT) technologies. This framework enables users to manipulate objects in both the physical and the virtual spaces through natural gestures and commands, leveraging AR for visual integration and IoT devices for physical state changes.

Table 12: Comparison of Zero-shot, Few-shot, and Chain of Thought outputs for a single example by GPT-4.

Model	Prompt	Aspect Prediction				Title Prediction
		Idea	Method	Outcome	Future	Title
GPT-3.5	0-Shot	0.869	0.859	0.873	0.881	0.896
	2-Shot	0.874	0.870	0.875	0.879	**0.913**
	CoT	0.835	0.850	0.857	0.864	0.893
	Few-Shot CoT	0.866	0.856	0.862	0.872	0.904
GPT-4	0-Shot	0.871	**0.872**	0.875	0.880	0.892
	2-Shot	0.872	0.870	0.875	0.874	0.910
	CoT	0.869	0.865	**0.877**	0.878	0.893
	Few-Shot CoT	0.869	0.865	0.874	0.869	0.902
Mistral 8x7B	0-Shot	0.869	0.869	0.875	0.881	0.884
	2-Shot	**0.876**	0.868	0.875	**0.882**	0.897
	CoT	0.857	0.866	0.858	0.870	0.884
	Few-Shot CoT	0.872	0.858	0.869	0.875	0.902

Table 13: Benchmark Results Measured by Cosine Similarity.

Model	Prompt	Aspect Prediction				Title Prediction
		Idea	Method	Outcome	Future	Title
GPT-3.5	0-Shot	0.839	0.845	0.855	0.860	0.875
	2-Shot	**0.872**	0.869	0.880	0.875	**0.892**
	CoT	0.860	0.858	0.852	0.875	0.870
	2-Shot CoT	0.867	**0.875**	0.878	**0.880**	0.891
GPT-4	0-Shot	0.815	0.783	0.812	0.814	0.869
	2-Shot	0.869	0.810	**0.886**	0.854	0.883
	CoT	0.829	0.806	0.841	0.837	0.869
	2-Shot CoT	0.868	0.858	0.880	0.863	0.884
Mistral 8x7B	0-Shot	0.823	0.822	0.840	0.838	0.822
	2-Shot	0.862	0.855	0.860	0.865	0.847
	CoT	0.829	0.821	0.839	0.850	0.828
	2-Shot CoT	0.870	0.866	0.875	0.877	0.862

Table 14: Benchmark Results Measured by BERTScore.

Model	Prompt	Aspect Prediction				Title Prediction
		Idea	Method	Outcome	Future	Title
GPT-3.5	0-Shot	0.014	0.017	0.032	0.027	0.068
	2-Shot	0.034	**0.029**	0.042	0.033	**0.101**
	CoT	0.015	0.018	0.020	0.023	0.050
	2-Shot CoT	0.026	0.025	0.031	0.027	0.079
GPT-4	0-Shot	0.008	0.006	0.012	0.009	0.049
	2-Shot	0.028	0.008	0.050	0.017	0.081
	CoT	0.010	0.007	0.021	0.013	0.052
	2-Shot CoT	0.025	0.019	0.041	0.016	0.064
Mistral 8x7B	0-Shot	0.014	0.014	0.027	0.020	0.020
	2-Shot	0.036	0.023	0.044	0.033	0.048
	CoT	0.014	0.014	0.023	0.020	0.011
	2-Shot CoT	**0.039**	0.026	**0.056**	**0.035**	0.060

Table 15: Benchmark Results Measured by BLEU.

Model	Prompt	Aspect Prediction				Title Prediction
		Idea	Method	Outcome	Future	Title
GPT-3.5	0-Shot	0.188	0.193	0.228	0.240	0.432
	2-Shot	0.275	**0.267**	0.287	0.276	**0.459**
	CoT	0.202	0.217	0.198	0.245	0.405
	2-Shot CoT	0.254	0.260	0.260	0.275	0.437
GPT-4	0-Shot	0.134	0.084	0.126	0.112	0.401
	2-Shot	0.269	0.138	**0.288**	0.210	0.436
	CoT	0.161	0.123	0.184	0.154	0.404
	2-Shot CoT	0.261	0.240	0.273	0.228	0.413
Mixtral 8x7B	0-Shot	0.173	0.168	0.206	0.179	0.287
	2-Shot	**0.288**	0.259	0.283	**0.279**	0.427
	CoT	0.170	0.164	0.202	0.206	0.275
	2-Shot CoT	0.286	0.264	**0.288**	0.293	0.436

Table 16: Benchmark Results Measured by ROUGE-1.

Model & Prompt	Key Idea	Method	Outcome	Projected Impact
GPT-4 0-Shot	0.454	0.383	0.390	0.411
GPT-4 2-Shot	0.414	0.373	0.424	0.439
GPT-4 0-CoT	0.360	0.348	0.377	0.407
GPT-4 2-CoT	0.386	0.358	0.436	0.425
GPT-3.5 0-Shot	0.402	0.378	0.398	0.425
GPT-3.5 2-Shot	0.387	0.377	0.398	0.414
GPT-3.5 0-CoT	0.330	0.357	0.365	0.352
GPT-3.5 2-CoT	0.372	0.382	0.390	0.391

Table 17: Downstream task evaluation results with human annotated summaries (BLEURT).

Model & Prompt	Key Idea	Method	Outcome	Projected Impact
GPT-4 0-Shot	**0.454**	**0.383**	0.390	0.411
GPT-4 2-Shot	0.414	0.373	0.424	**0.439**
GPT-4 0-CoT	0.360	0.348	0.377	0.407
GPT-4 2-CoT	0.386	0.358	**0.436**	0.425
GPT-3.5 0-Shot	0.402	0.378	0.398	0.425
GPT-3.5 2-Shot	0.387	0.377	0.398	0.414
GPT-3.5 0-CoT	0.330	0.357	0.365	0.352
GPT-3.5 2-CoT	0.372	0.382	0.390	0.391

Table 18: Downstream task evaluation results with human annotated summaries (FActScore).