Overview of the CLEF 2024 SimpleText Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications

Jennifer D'Souza[1,2], Salomon Kabongo[2], Hamed Babaei Giglou[1] and Yue Zhang[3]

[1]TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]Leibniz University of Hannover, Hannover, Germany
[3]Technische Universität Berlin, Germany

Abstract

This paper presents an overview of the CLEF 2024 SimpleText Task 4 on *SOTA? Tracking the State-of-the-Art in Scholarly Publications*, asking systems to perform two tasks: 1) classification – given the full text of an AI scientific paper, classify whether the paper indeed reports model scores on benchmark datasets, and if so, 2) information extraction – extract all pertinent (Task, Dataset, Metric, Score) tuples from the content of the scientific paper to automatically populate leaderboards used to keep track on the latest and greatest AI models. We discuss the details of the task set-up. First, the "SOTA?" task corpus comprising over 14K AI scientific papers, their corresponding annotations, and detailed corpus statistics. Second, the Evaluation Metrics used and the online Codalab Evaluation Platform to accept participant submissions. Third, the Results of the runs submitted by our participants.

Keywords

information extraction, leaderboards, benchmarks, artificial intelligence, open research knowledge graph, text mining, natural language processing, large language models

1. Introduction

This paper presents an overview of the CLEF 2024 SimpleText Task 4 on *SOTA? Tracking the State-of-the-Art in Scholarly Publications*, asking systems to perform two tasks: 1) classification – given the full text of an AI scientific paper, classify whether the paper indeed reports model scores on benchmark datasets, and if so, 2) information extraction – extract all pertinent (Task, Dataset, Metric, Score) tuples from the content of the scientific paper to automatically populate leaderboards used to keep track on the latest and greatest AI models. The task website is hosted at https://sites.google.com/view/simpletext-sota/. This task is a novel addition to the CLEF 2024 SimpleText Track [1]. It explores the structured scientific information model, as advocated by the Open Research Knowledge Graph (ORKG) project [2, 3], offering a new perspective on the objective of simplifying scientific information. Specifically, the task examines the phenomenon of leaderboards or scoreboards in Artificial Intelligence (AI) research. These leaderboards report new AI models and their performance in terms of the addressed tasks, evaluated datasets, and applied evaluation metrics.

The SimpleText track as a whole offers valuable data and benchmarks to facilitate discussions on the challenges associated with automatic text simplification. It presents an interconnected framework that encompasses various tasks, providing a comprehensive view of the complexities involved:

Task 1 on *Content Selection*: retrieve passages to include in a simplified summary.

Task 2 on *Complexity Spotting*: identify and explain diffficult concepts.

Task 3 on *Text Simplification*: simplify scientific text.

Task 4 on *SOTA?*: tracking the state-of-the-art in AI scholarly publications.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 9—12, 2024, Grenoble, France
jennifer.dsouza@tib.eu (J. D'Souza); kabenamualu@l3s.de (S. Kabongo)
https://sites.google.com/view/simpletext-sota (J. D'Souza)
0000-0002-6616-9509 (J. D'Souza); 0000-0002-0021-9729 (S. Kabongo); 0000-0003-3758-1454 (H. B. Giglou); 0009-0007-6432-1259 (Y. Zhang)

This paper presents an overview of the fourth task in the SimpleText track at CLEF 2024, i.e. tracking the state-of-the-art in scholarly publications. For a comprehensive overview of the other tasks, the task overview papers on Task 1 [4], Task 2 [5], and Task 3 [6], as well as the track overview paper [7], provide detailed information and further insights.

Aligned with the aim of simplifying scientific texts, is the goal of generating structured summaries of scientific knowledge [2] to enhance its machine-actionability. This entails making scholarly knowledge more amenable to advanced information technology tools, which, particularly in the face of the current proliferation of publications [8, 9], can significantly aid readers in monitoring scientific advancements. Related to the topic of monitoring scientific advancements is the concept of leaderboards in Artificial Intelligence (AI) research. Leaderboards are platforms that keep track of scores reported by various models introduced in the AI community in terms of certain integral elements: the models are evaluated on specific benchmark datasets, the datasets address a specific task, and the model output is evaluated by a performance metric [10]. This information is generally buried within the discourse text scholarly AI articles. Thus SimpleText in 2024 introduces a fourth task that handles the automatic text mining of the (Task, Dataset, Metric, Score) tuples from AI articles to automatically build leaderboards. This new task "***Task 4: SOTA?*** *Tracking the State-of-the-Art in Scholarly Publications*" is explained in detail in section 2.

A total of 45 teams registered for our SimpleText track at CLEF 2024. A total of 20 teams submitted 207 runs in total for the Track, of which 2 teams submitted a total of 36 runs for Task 4.

The rest of the paper is organized as follows: Section 2 outlines the overall task definition and objectives (2.1), followed by an in-depth exploration of the task dataset and its statistics (2.2). This section also discusses the use of the Codalab competition site for accepting participant submissions (2.3) and concludes with the evaluation metrics used (2.4). Section 3 provides an overview of the classification and information extraction methods adopted by the two teams. In Section 4, we present and discuss the results of the official submissions. Finally, Section 5 concludes the paper with a summary of the findings.

2. Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications

This section details *Task 4: SOTA?* on tracking the state-of-the-art in scholarly publications.

2.1. Description

In Artificial Intelligence (AI), a common research objective is the development of new models that can report state-of-the-art (SOTA) performance. The reporting usually comprises four integral elements: Task, Dataset, Metric, and Score. These (Task, Dataset, Metric, Score) tuples or (T, D, M, S) hence, coming from various AI research papers go on to power leaderboards in the community. Leaderboards, akin to scoreboards, traditionally curated by the community, are platforms displaying various AI model scores for specific tasks, datasets, and metrics. Examples of such platforms include the benchmarks feature on the Open Research Knowledge Graph and Papers with Code (PwC). Utilizing text mining techniques allows for a transition from the conventional community-based leaderboard curation to an automated text mining approach. Consequently, the goal of Task 4: SOTA? is to develop systems that can classify whether a scholarly article provided as input to the model reports a (T, D, M, S) or not. And for articles reporting (T, D, M, S), extract all the relevant ones from the paper text.

The Task 4: SOTA? task formalism is defined as follows: given the text of a scientific paper A, the goal is to extract its Leaderboards L, where $L = \{l_1, ..., l_x\}$ and A can have between one to an undefined number of Leaderboards. Each Leaderboard l comprises the (T, D, M, S) quadruple.

This task was divided into two separate evaluation phases:

Evaluation Phase I. *Few-shot (T,D,M,S) extraction:* Systems are expected to identify whether an incoming AI paper reports leaderboards or not; and for paper's reporting leaderboards, extract all the

pertinent (T, D, M, S) quadruples. The "few-shot" aspect of this subtask is that it involves (T, D, M) labels previously seen in the training dataset.
Evaluation Phase II. *Zero-shot (T,D,M,S) extraction:* This is similar to Subtask 4.1, but involves a new test dataset containing (T, D, M) tuples that were not seen in the training set, testing the system's ability to handle zero-shot scenarios.

2.2. Dataset

The training and test datasets for *Task 4: SOTA?* were derived from the community-curated (T, D, M, S) annotations of thousands of AI articles available on https://paperswithcode.com/ (PwC) (CC BY-SA). We used the dataset from our prior work, specifically the PwC data downloaded on December 09, 2023 [11]. The corpus included over 8,000 articles, with 7,987 used for training and 994 for testing, divided into 751 for the few-shot setting and 241 for the zero-shot setting. While the annotations came from PwC, the full-text of the articles was sourced from the arXiv preprint server under CC-BY licenses. Each article in the dataset is available in TEI XML format, complete with one or more (T, D, M, S) annotations from PwC. The complete *Task 4: SOTA?* dataset is publicly released on GitHub under the CC-BY-SA 4.0 license, accessible at https://github.com/jd-coderepos/sota.

Another important subset of our dataset, in addition to corpus with (T, D, M, S) annotations, was the "no leaderboards papers" i.e. compiling a set of AI papers that did not report leaderboards. We included a set of approximately 4,401 and 648 articles that do not report leaderboards into the train and test sets, respectively. These articles were randomly selected by leveraging the arxiv category feature, then filtering it to papers belonging to domains unrelated to AI/ML/Stats. These articles were annotated with the *unanswerable* label. Thus given the overall dataset, systems could perform the expected task i.e. classification and information extraction.

2.2.1. Dataset Statistics

We now provide detailed statistics of our corpus, focusing on the granularity of different annotation counts and coverage of the annotation labels.

Overall, the train and validation datasets contained 12,288 and 100 papers, respectively. The train dataset included 7,936 papers annotated with leaderboards, while the remaining 4,352 papers lacked such annotations and were marked as "unanswerable." The validation dataset featured 51 papers with leaderboard annotations and 49 without. In the few-shot test dataset for evaluation phase 1, there were 1,401 papers, split between 753 with leaderboards and 648 marked as "unanswerable." The zero-shot test dataset for evaluation phase 2 comprised 789 papers, with 241 having leaderboard annotations and 548 labeled as "unanswerable."

Table 1 shows the counts of the unique (Task, Dataset, Metric) entities or elements in the Task 4: SOTA? dataset.[1] The novelty is most pronounced in Datasets, followed by Metrics, and then Tasks. This novelty partially stems from the community-curated annotations in the PwC, which result in unnormalized labels. For instance, the metric "F1-score" might be recorded as "F1," "F-score," or "F-measure," and each variation is considered a unique Metric label. This diversity aims to mirror the variability seen in scientific papers, where, to our knowledge, there is no standardized naming convention for these entities. As *SOTA?* focuses on information extraction, we intend for the variety of community-curated annotations to reflect the terminology used in the source papers. However, this diversity might also mirror the annotators' preferences within the PwC, and the annotated dataset did not guarantee uniformity.

Tables 2 and 3 display the top 10 most frequent (Task, Dataset, Metric) annotations in the *SOTA?* dataset, both as individual elements and as combined triples. This may also indicate a prevailing research trend within the scientific community: "Image Classification" is a commonly addressed task, and the

[1]Since the Score element varies continuously, we do not consider counting unique occurrences as a valid statistic for the Task 4: SOTA? dataset.

Table 1
SimpleText Task 4: SOTA? dataset statistics displaying unique labels for annotated (Task, Dataset, Metric) elements.

Parameter	Train	Few-shot Test	Zero-shot Test
Unique Tasks	1,372	320	236
Unique Datasets	4,795	935	646
Unique Metrics	2,782	637	397
Unique (Task, Dataset, Metric) triples	11,977	1,900	1,262
Avg. (Task, Dataset, Metric) triples per paper	6.93	5.69	7.53

"ImageNet" dataset is frequently used to develop or evaluate systems, often employing variants of the "accuracy" metric.

Table 2
Ten most common Tasks, Datasets, and Metrics in the SimpleText Task 4: SOTA? training dataset.

Task	Frequency	*Dataset*	Frequency	*Metric*	Frequency
Image Classification	2,273	ImageNet	1,603	Accuracy	4,383
Atari Games	1,448	COCO Test-Dev	792	Score	1,515
Node Classification	1,113	Human3.6M	624	F1	1,384
Object Detection	1,001	CIFAR-10	585	PSNR	1,144
Video Retrieval	997	COCO Minival	310	MAP	1,068
Link Prediction	941	YouTube-VOS 2018	295	MIoU	862
Semantic Segmentation	901	CIFAR-100	252	SSIM	799
Semi-supervised Video Object Segmentation	890	MSR-VTT-1kA	247	Top 1 Accuracy	789
3D Human Pose Estimation	889	FB15k-237	244	1:1 Accuracy	787
Question Answering	866	MSU Super-Resolution for Video Compression	225	Number of Params	759

Table 3
Ten most common (Task, Dataset, Metric) triples in the SimpleText Task 4: SOTA? training dataset.

(*Task*, *Dataset*, *Metric*)	Frequency
(Image classification, ImageNet, Top 1 accuracy)	524
(Image classification, ImageNet, Number of params)	313
(Image classification, ImageNet, GFLOPs)	256
(3D human pose estimation, Human3.6M, Average MPJPE)	197
(Image classification, CIFAR-10, Percentage correct)	128
(Action classification, Kinetics-400, ACC@1)	108
(Object detection, COCO test-dev, Box mAP)	106
(Image classification, CIFAR-100, Percentage correct)	105
(Semantic segmentation, ADE20K, Validation mIoU)	92
(Neural architecture search, ImageNet, Top-1 error)	83

As alluded to earlier, the *Task 4: SOTA?* dataset does not guarantee that the community curated PwC (T, D, M, S) annotations for each respective paper matches with the text used in the source scientific article. With the statistics shown in Table 4, we offer insights to what extent of in what proportion of the annotations, the respective (T, D, M, S) labels were found in the underlying source text across the Train and the two Test datasets. In the training dataset, the occurrence of annotation labels in the accompanying paper's full text varies by category: 60.24% for Tasks, 58.86% for Scores, 45.48% for

Datasets, and 42.69% for Metrics. This data indicates that Metrics exhibit the greatest inconsistency in annotation labels, followed by Datasets, Scores, and Tasks. Similar patterns were reflected in the Test datasets. We offer the reader this perspective in interpreting the performance scores obtained by the two participants in this task–this year's Task 4: SOTA? dataset presents the most variability in annotations in the training and evaluation of participant systems which in turn can account for lower reported scores.

Table 4
SimpleText Task 4: SOTA? dataset statistics showing the proportion of annotated elements (Task, Dataset, Metric, Score), where the annotation label text exactly matches the text found within the paper.

Dataset Count Parameter	Train	Few-shot Test	Zero-shot Test
Unique *Tasks* per Paper	10,810	1,008	351
Unique found-in-paper *Tasks* per Paper	6,512	649	222
Ratio *Tasks*	0.6024	0.6438	0.6325
Unique *Datasets* per Paper	21,278	1,937	777
Unique found-in-paper *Datasets* per Paper	9,677	816	328
Ratio *Datasets*	0.4548	0.4213	0.4221
Unique *Metrics* per Paper	23,220	2,136	702
Unique found-in-paper *Metrics* per Paper	9,913	861	340
Ratio *Metrics*	0.4269	0.4031	0.4843
Unique *Scores* per Paper	52,092	4,110	1,688
Unique found-in-paper *Scores* per Paper	30,660	2,266	911
Ratio *Scores*	0.5886	0.5513	0.5462

To support future research and evaluations, the "Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications" task dataset is now publicly available on GitHub under a CC BY-SA 4.0 license. It can be accessed at https://github.com/jd-coderepos/sota/.

2.3. Codalab Competition Site

Automated evaluations for the participants' systems were implemented via the Codalab competitions' website [12]. CodaLab is well-known for hosting a variety of machine learning and data science competitions. It provides a comprehensive environment where competition organizers can manage entries, participants can submit solutions, and results can be evaluated automatically based on predefined criteria.

2.3.1. Setup

To configure *Task 4: SOTA?* on CodaLab, we followed the official step-by-step guide guide. In the main configuration file, COMPETITION.YAML, we outlined the competition's evaluation phases as follows:

1. **Practice Phase:** This initial phase allows participants to make submissions on the development set, testing for valid file formats and verifying that the system returns the expected scores.
2. **Evaluation Phases:** We established two main evaluation phases for the task:
    - **Evaluation Phase 1:** Starting at midnight on April 23, 2024, and transitioning directly into Evaluation Phase 2.
    - **Evaluation Phase 2:** Beginning at midnight on April 29, 2024, and concluding at midnight on May 4, 2024.

During each phase, competition organizers can upload the respective phase's test dataset annotations to CodaLab and assign them to the designated phase. The test dataset annotations remain hidden from the participants. Additionally, a custom scoring program, written in Python, can also be associated with each evaluation phase. Participants then for the current running evaluation phase can upload their system output in the designated format prescribed by the competition organizers. We detail the *SOTA?* submission format next.

2.3.2. Submission Format

Participants were required to generate a submission folder containing the annotations in a specified output format, based on the input data provided. The input data consisted of a main folder with several subfolders, each labeled with a unique article ID. Each subfolder included the full text of an AI paper in "tei.xml" format. Participants were to apply their systems to this dataset and produce an output folder mirroring the input structure, with identically numbered subfolders to ensure each article could be uniquely identified during evaluation. Each output subfolder was expected to contain a file named "annotations.json," which either contained (T, D, M, S) annotations or the string "unanswerable" if the system determined that the input paper did not report a leaderboard.

(T, D, M, S) annotations format. For papers with leaderboards, their annotations were expected in a JSON file per the format shown in Figure 1.

Our competition site is now live and can be accessed at https://codalab.lisn.upsaclay.fr/competitions/16616. Although the two official evaluation phases have concluded, the site will continue to operate indefinitely in a post-competition phase, hosting the zero-shot evaluation dataset.

2.4. Evaluation Metrics

We conducted three main categories of evaluations.

1. **Classification Accuracy**: This metric measured the accuracy with which the participant systems identified the "unanswerable" papers i.e. papers without leaderboards compared with the gold-standard.
2. **Summarization Rouge**: The ROUGE metrics [14] are commonly used for evaluating the quality of text summarization systems. ROUGE-1 measures the overlap of unigram (single word) units between the generated summary and the reference summary. ROUGE-2 extends this to measure the overlap of bigram (two consecutive word) units. ROUGE-L calculates the longest common subsequence between the generated and reference summaries, which takes into account the order of words. ROUGE-LSum is an extension of ROUGE-L that considers multiple reference summaries by treating them as a single summary. These metrics provide a quantitative assessment of the similarity between the generated and reference summaries, helping researchers and developers evaluate and compare the effectiveness of different summarization approaches. They have become widely used benchmarks in the field of automatic summarization. We treated the (T, D, M, S) extraction task as analogous to a summarization objective and hence reported system overall extraction performance based on the ROUGE summarization metrics.
3. **Per (T, D, M, S) Element-wise Extraction F1-score**: In this evaluation category, we evaluated the model JSON output in a fine-grained manner w.r.t. each of the individual (T, D, M, S) elements and overall for which we reported the results in terms of the standard recall, precision, and F1 score. In addition, these element-wise extraction evaluation results are reported using exact match of the extracted strings with the gold-standard as well as using partial match between the

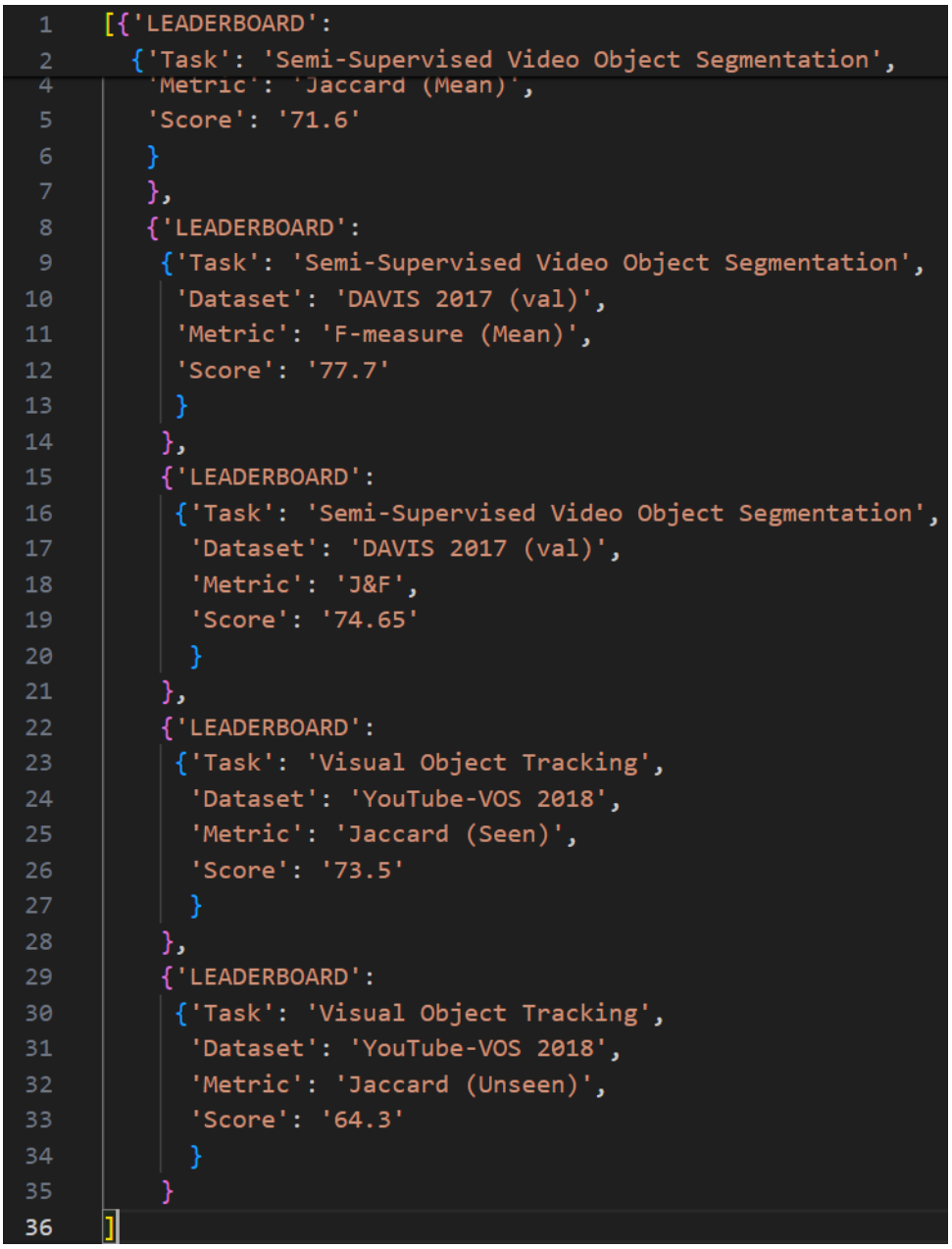

Figure 1: Submission format example for one paper containing (T, D, M, S) annotations. This file is publicly released online and shows the leaderboard annotations for the paper titled "Proposal, tracking and segmentation (pts): A cascaded network for video object segmentation" [13].

extracted string and the gold-standard. The partial matches were computed using https://github.com/seatgeek/thefuzz python library. Referring back to Figure 1 for the expected (T, D, M, S) annotation format, the evaluation script was written to handle the fact that predicted leaderboards could exceed or be less than the total number of gold-standard leaderboards. Additionally that the order of (T, D, M, S) leaderboards in the system predictions was not expected to match the order of the (T, D, M, S) leaderboards in the gold-standard since there was no precendence format prespecified. To handle this, the script operates in two steps: it first compares each predicted (T, D, M, S) unit to the gold standard to find the best match, and then it calculates the individual element-wise extraction measures to determine the overall system recall, precision, and F1-score.

The official SimpleText Task 4: SOTA? evaluation script is publicly released online.

3. Task 4: SOTA? Participant Approaches

In this section, we discuss participant submissions to "Task 4: SOTA? Tracking the State-of-the-Art in Scholarly Publications" task of the SimpleText track at CLEF 2024. A total of 2 teams submitted 36 runs in total.

AMATU by Staudinger et al. [15] submitted a total of three runs for the **few-shot** evaluation phase of Task 4. They submitted nine runs for the **zero-shot** evaluation phase of Task 4. Their general approach to extract the (T, D, M, S) annotations were in two main categories: 1) a pure pattern-based approach inspired after AxCell [16], and 2) an AI-based approach using LLMs with a zero-shot prompt and a few-shot prompt tested for GPT-3.5 [17] and Mistral-7B [18]. For the latter category, they also experimented with variants on the input scholarly article text from which the (T, D, M, S) annotations were expected to be extracted. This we generally refer to as the *context*. They tried two context variants: 1) full paper text and 2) only the text from sections referring to experiments and results, in addition to the abstract, which was pre-extracted inspired by the Argumentative Zoning (AZ) method [19].

L3S by Kabongo et al. [20] submitted a total of 12 runs for the **few-shot** evaluation phase of Task 4. They submitted 12 runs for the **zero-shot** evaluation phase of Task 4. Their approach entailed leveraging the FLAN-T5 [21] strategy which encompassed fine-tuning a pre-trained LLM with a standard set of instructions to better equip them to handle various tasks. Leveraging the applicable instructions from the FLAN-T5 collection, they fine-tuned LLMs, viz. Mistral-7B and LLaMA 2 [22], to make them better suited to handle the (T, D, M, S) extraction task. Furthermore, they also tested the most recent proprietary GPT models viz. GPT-4 [23] and GPT-4o. Finally, as the information extraction context they tried 3 different methods: DocTAET ((T)-title, (A)- abstract, (E)-experimental setup, and (T)-tabular information parts of the full-text), DocREC (text selected from the sections named (R)-results, (E)-experiments, and (C)-conclusions), and DocFULL (full paper text). Resultingly, for each evaluation phase they submitted a total of 4 models x 3 contexts = 12 runs.

We encourage readers to refer to the referenced participant papers for detailed explorations of their approaches and the motivations behind their systems.

4. Task 4: SOTA? Results and Discussion

Tables 5 and 6 present a summary of the results from the two teams. Overall, given *Team AMATU*'s approaches, the pattern-based method proved a competitive solution to the SOTA challenge, in comparison to advanced LLM-based solutions. While the LLM solution did outperform the pattern-based approach the difference was minor. Furthermore, comparing the zero-shot versus few-shot paradigms, the LLMs were significantly more effective in the few-shot setting i.e. when shown successful task completion outputs. Also, the LLM performed significantly better when given the full paper text as input from which to extract (T, D, M, S) as opposed to given selective text using the AZ method.

For *Team L3S*, in both the evaluation phases, their model results showed that minimal finetuning of relatively smaller LLMs, specifically Mistral-7B, equips them for (T, D, M, S) extraction task surpassing the performance of LLMs, specifically the latest GPT-4 proprietary models, with a significantly more vast parameter space. The overall best results even for the extraction of the (T, D, M, S) elements was obtained by Mistral given the DocTAET context.

Comparing *Team AMATU* and *Team L3S*, none of the systems from the former team were finetuned

to the task. Thus *Team AMATU* presents novel insights into the community to leveraging LLM's effectively for the (T, D, M, S) extraction objective using clever prompt engineering strategies that shows comparable performans to the latter teams' computationally intensive finetuning approach. It maybe that finetuning would be essential to create the most optimal model, however, from the team's solutions the importance prompt engineering for effective downstream performance is clearly emphasized.

Table 5
Evaluation results for the binary classification or filtering of papers with and without leaderboards (reported as General Accuracy) and as a structured summary generation task (reported with ROUGE metrics). *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

	Few-shot					**Zero-shot**				
	Rouge				**Gen.**	**Rouge**				**Gen.**
	1	**2**	**L**	**Lsum**	**Acc.**	**1**	**2**	**L**	**Lsum**	**Acc.**
AMATU	**58.34**	12.98	**57.34**	54.4	75.59	**73.72**	6.07	72.72	72.57	85.93
L3S	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	**73.01**	**72.95**	**95.97**

Table 6
Evaluation results w.r.t. the individual (Task, Dataset, Metric, Score) elements and Overall in terms of **F1 score**. *Team AMATU*'s few-shot evaluation results are reported for AxCell and their zero-shot evaluation results are reported for GPT-3.5 via the few-shot prompting paradigm. *Team L3S*'s results are reported for Mistral-7B finetuned with the DocTAET context. The best results are shown in bold.

		Few-shot					**Zero-shot**				
Model	**Mode**	**T**	**D**	**M**	**S**	**Overall**	**T**	**D**	**M**	**S**	**Overall**
AMATU	Exact	27.11	**23.22**	**24.85**	**9.34**	**21.13**	10.01	13.16	11.65	**9.85**	11.16
	Partial	28.08	24.92	25.8	10.86	22.62	16.12	17.12	13.72	11.1	14.52
L3S	Exact	**33.38**	18.51	24.23	1.87	19.50	**26.99**	**14.32**	**22.04**	1.20	**16.14**
	Partial	46.35	32.75	34.16	2.25	28.88	44.90	27.29	32.23	1.41	26.46

5. Conclusions

This concludes the results for the CLEF 2024 SimpleText Task 4: SOTA? on tracking the state-of-the-art in scholarly publications. Our main findings are the following: First, effective prompting paradigms should be a go-to strategy to test LLMs out-of-the-box for the SOTA? shared task objective. Second, finetuning small-scale models makes them better able to handle the SOTA? objective than larger-scale LLMs known for their generative AI abilities when simply applied to the IE task. Third, the paper context over which the IE task is expected to be performed must have an ideal balance of length versus selectivity of specific sections in the paper that indeed are highly likely to contain mentions of the (T, D, M, S). On the extreme end of the spectrum, using the full paper text without effective context selection hinders and seems to distract the LLM downstream IE task performance.

Acknowledgments

The "SOTA?" track as Task 4 within the SimpleText 2024 evaluation lab at CLEF 2024 has been jointly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number: NFDI4DataScience (460234259) and the German BMBF project SCINEXT (01lS22070).

References

[1] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. Di Nunzio, F. Vezzani, J. D'Souza, S. Kabongo, H. B. Giglou, Y. Zhang, S. Auer, J. Kamps, Clef 2024 simpletext track: Improving access to scientific texts for everyone, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part VI, Springer-Verlag, Berlin, Heidelberg, 2024, p. 28–35. URL: https://doi.org/10.1007/978-3-031-56072-9_4. doi:10.1007/978-3-031-56072-9_4.

[2] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. E. Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, Bibliothek Forschung und Praxis 44 (2020) 516–529.

[3] J. D'Souza, S. Kabongo, M. Prinz, Y. Jaradeh, K. E. Farfar, Orkg benchmarks, in: S. Auer, V. Ilangovan, M. Stocker, S. Tiwari, L. Vogt (Eds.), Open Research Knowledge Graph, Cuvillier, Göttingen, Germany, 2024, pp. 49–56.

[4] E. SanJuan, S. Huet, J. Kamps, L. Ermakova, Overview of the clef 2024 simpletext task 1: Retrieve passages to include in a simplified summary, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.

[5] L. Ermakova, H. Azarbonyad, S. Bertin, O. Augereau, Overview of the CLEF 2023 SimpleText Task 2: Difficult Concept Identification and Explanation, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.

[6] L. Ermakova, S. Bertin, H. McCombie, J. Kamps, Overview of the CLEF 2023 SimpleText Task 3: Scientific text simplification, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.

[7] L. Ermakova, E. SanJuan, S. Huet, H. Azarbonyad, G. M. D. Nunzio, F. Vezzani, J. D'Souza, J. Kamps, Overview of the CLEF 2024 simpletext track — improving access to scientific texts for everyone, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany, 2024.

[8] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., Science of science, Science 359 (2018) eaao0185.

[9] L. Bornmann, R. Haunschild, R. Mutz, Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases, Humanities and Social Sciences Communications 8 (2021) 1–15.

[10] Y. Hou, C. Jochim, M. Gleize, F. Bonin, D. Ganguly, Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019. URL: https://aclanthology.org/P19-1513. doi:10.18653/v1/P19-1513.

[11] S. Kabongo Kabenamualu, J. D'Souza, S. Auer, Effective context selection in llm-based leaderboard generation: An empirical study, in: Proceedings of the 29th International Conference on Natural Language & Information Systems, 2024. URL: https://nldb2024.di.unito.it/.

[12] A. Pavao, I. Guyon, A.-C. Letournel, D.-T. Tran, X. Baro, H. J. Escalante, S. Escalera, T. Thomas, Z. Xu, Codalab competitions: An open source platform to organize scientific challenges, Journal of Machine Learning Research 24 (2023) 1–6. URL: http://jmlr.org/papers/v24/21-1436.html.

[13] Q. Zhou, Z. Huang, L. Huang, Y. Gong, H. Shen, C. Huang, W. Liu, X. Wang, Proposal, tracking and segmentation (pts): A cascaded network for video object segmentation, arXiv preprint arXiv:1907.01203 (2019).

[14] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization

branches out, 2004, pp. 74–81.
[15] M. Staudinger, A. El-Ebshihy, A. M. Ningtyas, F. Piroi, A. Hanbury, AMATU@Simpletext2024: Are LLMs alone any good for Scientific Entity Extraction?, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.
[16] M. Kardas, P. Czapla, P. Stenetorp, S. Ruder, S. Riedel, R. Taylor, R. Stojnic, Axcell: Automatic extraction of results from machine learning papers, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 8580–8594.
[17] OpenAI, Gpt-3.5 turbo documentation, 2023. URL: https://platform.openai.com/docs/models/gpt-3-5-turbo, accessed: 2024-06-10.
[18] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[19] S. Teufel, et al., Argumentative zoning: Information extraction from scientific text, Ph.D. thesis, Citeseer, 1999.
[20] S. Kabongo, J. D'Souza, S. Auer, Exploring the latest llms for leaderboard extraction, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, Online, 2024.
[21] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, Journal of Machine Learning Research 25 (2024) 1–53.
[22] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al., Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288 (2023).
[23] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).