Exploring the Latest LLMs for Leaderboard Extraction

Leibniz University of Hannover, Germany

Authors : Salomon Kabongo, Dr. Jennifer D'Souza and Prof. Sören Auer

SimpleText @ Task 4: SOTA ?

Agenda

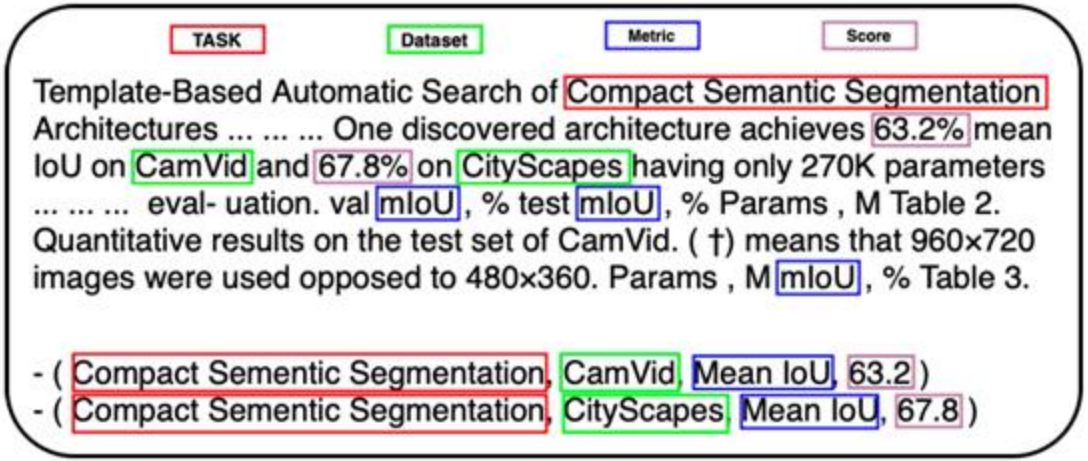
- Background
- Definition
- Our approach
- The Task Corpus
- Results
- Conclusion

Background

- The rapid advancements in Large Language Models (LLMs) have opened new avenues for automating complex tasks in AI research.
- This work investigates the efficacy of different LLMs-Mistral 7B, Llama-2, GPT-4-Turbo and GPT-4.o in extracting leaderboards from AI research articles.
- We explored three types of contextual inputs to the models: **DocTAET** (Document Title, Abstract, Experimental Setup, and Tabular Information), **DocREC** (Results, Experiments, and Conclusions), and **DocFULL** (entire document).
- We evaluates the performance of these models in generating **(Task, Dataset, Metric, Score)** quadruples from papers. The findings reveal significant insights into the strengths and limitations of each model and context type.

Background

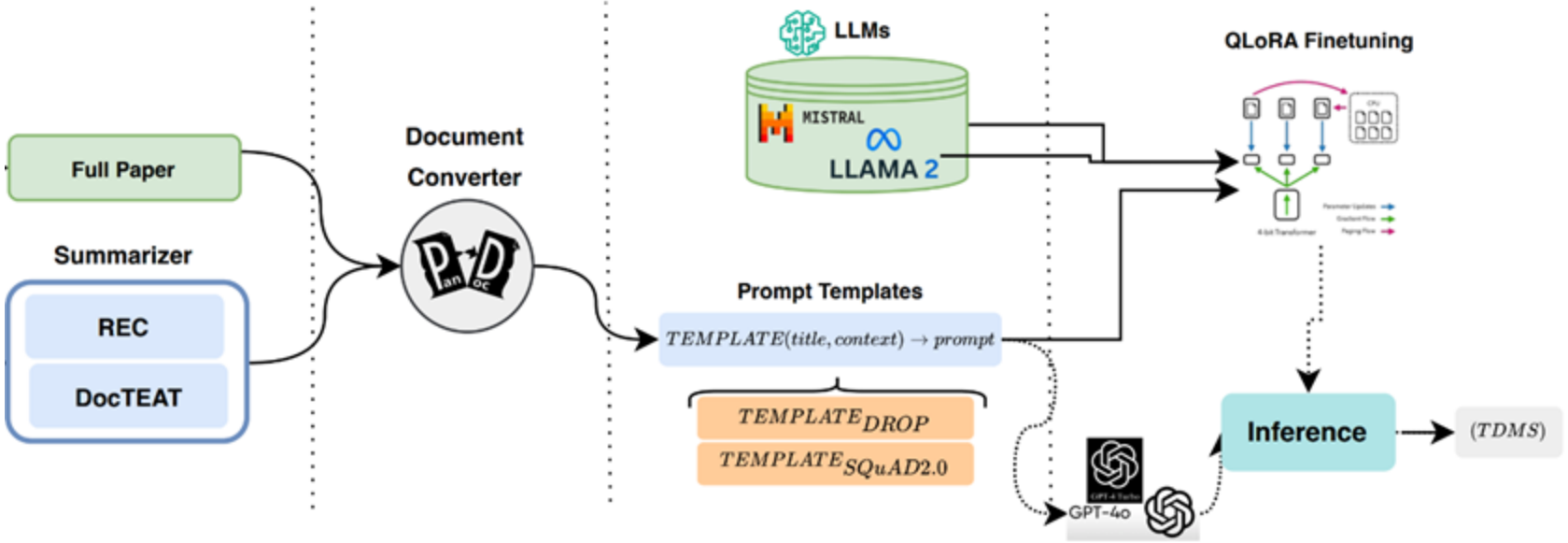
- Empirical Machine learning studies **how machines learn** with respect to **a task**, a **performance metric**, and **a dataset** (Mitchell, 2006). The **(task, dataset, metric name, metric value)** tuple can therefore be seen as representing a single result (Leaderboard) of a machine learning paper.



Definition

- **DocTAET**: comprises text selected from the (T)-title, (A)-abstract, (E)-experimental setup, and (T)- tabular information parts of the full-text. It yields an average context length of *493 words*.
- **DocREC**: Introduced for the first time in this work, the DocREC context comprises text selected from the sections named (R)-results, (E)-experiments, and (C)-conclusions. It yields an average context length of *1,586 words.*
- **DocFULL**: we used the full paper text as context. This approach entailed compiling the LaTeX source code of the document and translating its entirety into a plain text file. It yields an average context length of *5,948 words*.

Our Approach



The Task Corpus

Table 1
Our DocREC (Documents Result[s], Experimentation[s] and Conclusion) corpora statistics. The "papers w/o leaderboard" refers to papers that do not report leaderboard.

	Our Corpus		
	Train	Test-Few-shot	Test Zero-shot
Papers w/ leaderboards	7,987	753	241
Papers w/o leaderboards	4,401	648	548
Total TDM-triples	415,788	34,799	14,800
Distinct TDM-triples	11,998	1,917	1,267
Distinct *Tasks*	1,374	322	236
Distinct *Datasets*	4,816	947	647
Distinct *Metrics*	2,876	654	412
Avg. no. of TDM per paper	5.12	4.81	6.11
Avg. no. of TDMS per paper	6.95	5.81	7.86

Results - Summarization

Table 2

Evaluation results of Llama-2, Mistral, GPT-4-Turbo, and GPT-4.0 for the shared task, reported using the metrics proposed for the task. The output evaluations are conducted as a structured summary generation task (reported with ROUGE metrics) and as a binary classification task to distinguish between papers with and without leaderboards (reported as General Accuracy).

	Few-shot					Zero-shot					
Model	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**	**General -Accuracy**	**Rouge1**	**Rouge2**	**RougeL**	**RougeLsum**	**General -Accuracy**	**Context**
Llama-2 7B	49.68	10.18	48.91	49.02	83.51	68.15	4.81	67.59	67.78	86.82	DocREC
	49.70	17.62	48.81	48.81	83.62	62.75	10.88	62.07	62.18	86.22	DocTAET
	5.38	0.79	4.96	5.13	57.54	7.55	0.71	7.24	7.35	37.80	DocFULL
Mistral 7B	55.46	14.11	54.54	54.64	88.44	72.98	6.87	72.42	72.35	92.40	DocREC
	57.24	**19.67**	56.28	**56.19**	**89.68**	73.54	**12.23**	73.01	72.95	**95.97**	DocTAET
	6.73	0.77	6.36	6.49	58.45	9.38	0.59	9.11	9.23	39.28	DocFULL
GPT-4-Turbo	52.64	5.82	51.99	51.76	60.89	72.80	2.66	72.35	72.09	77.06	DocREC
	43.14	2.41	42.97	42.91	47.33	59.98	0.48	59.89	59.74	61.18	DocTAET
	48.50	3.21	48.06	47.96	52.87	70.10	1.8	69.75	69.73	72.65	DocFULL
GPT-4.o	**58.59**	16.81	**56.37**	55.45	83.21	**74.94**	9.02	**73.65**	**73.02**	87.94	DocREC
	52.10	13.72	50.77	49.26	80.63	69.59	8.81	68.57	67.43	87.56	DocTAET
	55.41	17.82	53.01	51.79	79.56	70.05	10.89	68.42	67.51	78.95	DocFULL

Results - F1 Score

Table 3

Evaluation results of Llama-2, Mistral, GPT-4-Turbo, and GPT-4.0 for the shared task, reported using the metrics proposed for the task. The evaluation considers the individual (Task, Dataset, Metric, Score) elements and Overall in the model JSON generated output, reported in terms of **F1 score**.

		Few-shot					Zero-shot					
Model	**Mode**	**Task**	**Dataset**	**Metric**	**Score**	**Overall**	**Task**	**Dataset**	**Metric**	**Score**	**Overall**	**Context**
Llama-2 7B	Exact	20.93	13.06	13.96	3.04	12.75	13.97	6.83	11.72	2.61	8.78	DocREC
	Partial	31.37	22.50	21.99	3.46	19.83	24.05	16.6	18.28	3.10	15.51	
	Exact	29.53	16.68	20.02	1.14	16.84	21.75	11.26	16.99	0.77	12.69	DocTAET
	Partial	43.37	30.36	30.51	1.38	26.40	38.48	23.10	27.09	0.96	22.41	
	Exact	1.59	1.36	0.94	0.23	1.03	2.06	1.30	1.52	0.33	1.30	DocFULL
	Partial	2.29	1.82	1.68	0.37	1.54	3.36	2.49	2.49	0.54	2.22	
Mistral 7B	Exact	26.77	15.68	18.70	6.36	16.88	17.99	11.80	15.55	5.04	12.60	DocREC
	Partial	39.75	27.28	28.49	7.08	25.65	29.88	21.05	23.16	5.75	19.96	
	Exact	**33.38**	**18.51**	**24.23**	1.87	**19.50**	**26.99**	14.32	**22.04**	1.20	**16.14**	DocTAET
	Partial	**46.35**	**32.75**	**34.16**	2.25	28.88	**44.90**	27.29	**32.23**	1.41	26.46	
	Exact	0.81	0.57	0.57	0.56	0.63	0.22	0.33	0.33	0.76	0.42	DocFULL
	Partial	1.19	0.85	0.81	0.84	0.92	0.56	0.67	0.78	0.87	0.72	
GPT-4-Turbo	Exact	7.61	6.19	4.92	4.25	5.74	4.26	5.35	3.86	3.28	4.18	DocREC
	Partial	16.48	13.96	11.03	7.03	12.13	13.76	11.09	10.19	5.46	10.13	
	Exact	2.99	2.69	0.95	0.75	1.84	1.13	0.79	0.34	0.11	0.59	DocTAET
	Partial	6.22	5.42	3.03	1.63	4.08	2.72	1.59	1.59	0.11	1.5	
	Exact	3.38	3.16	1.98	2.48	2.75	2.45	2.98	1.81	2.77	2.5	DocFULL
	Partial	7.03	6.41	4.96	4.15	5.64	6.49	5.85	4.47	3.56	5.09	
GPT-4.o	Exact	16.14	16.11	15.50	10.76	14.63	16.04	**15.05**	17.43	10.38	14.72	DocREC
	Partial	38.40	32.63	29.35	15.20	**28.90**	37.23	**31.16**	29.97	**14.96**	**28.33**	
	Exact	14.10	12.76	9.91	2.11	9.72	13.78	10.25	11.01	2.36	9.35	DocTAET
	Partial	31.84	26.65	20.83	4.22	20.92	29.33	23.87	19.50	3.71	19.12	
	Exact	16.72	14.53	14.67	**11.25**	14.29	13.08	14.94	16.09	**11.17**	13.82	DocFULL
	Partial	36.56	31.0	27.61	**16.50**	27.93	35.59	28.28	27.38	14.80	26.52	

Conclusion

- Our participation in the shared task has demonstrated that fine-tuning open-source models like Mistral 7B and Llama-2 7B can yield competitive, and in some cases superior, results compared to proprietary models such as GPT-4.o and GPT-4-Turbo.
- Throughout our experiments, the DocTAET context typically delivered dependable and accurate performance, while the DocREC context excelled in scenarios where precision is paramount.
- In conclusion, our involvement in the shared task has not only highlighted the effectiveness of fine-tuned open-source models but also emphasized the importance of strategic context selection in maximizing model performance.

Question ?

Thank you for your attention

By Salomon Kabongo, Dr. Jennifer D'Souza and Prof. Sören Auer

Presented by Salomon Kabongo

L3S Research Center, Leibniz University of Hannover, TIB, Germany

SimpleText @ Task 4: SOTA ?