AirQA: A Comprehensive QA Dataset for AI Research with Instance-Level Evaluation

Tiancheng Huang, Ruisheng Cao, Yuxin Zhang, Zhangyi Kang, Zijian Wang, Chenrun Wang, Yijie Luo, Hang Zheng, Lirong Qian, Lu Chen, Kai Yu

Motivation

Dataset	# QA	Evaluation Methods	Task types				Question based on				
			Sgl.	Multi.	Retr.	Comp.	Full Text	Table	Image	Form.	Meta.
ScholarlyRead [30]	10K	BLEU, METEOR, ROUGE	✓	✗	✗	✗	✗	✗	✗	✗	✗
QASPER [9]	5,049	F1	✓	✗	✗	✗	✗	✗	✗	✗	✗
QASA [21]	1,798	Precision, Recall, F1, ROUGE	✓	✗	✗	✗	✓	✗	✗	✗	✗
SPIQA [28]	270K	METEOR, CIDEr, ROUGE, BERTScore, LLMLogScore	✓	✗	✗	✗	✓	✓	✓	✗	✗
PeerQA [5]	579	MRR, Recall, Rouge-L, AlignScore, Prometheus-2	✓	✗	✗	✗	✓	✗	✗	✓	✗
SciDQA [32]	2,937	ROUGE, BLEURT-20, BERTScore, LLM judge	✓	✓	✗	✗	✓	✓	✓	✓	✗
M3SciQA [22]	1,452	MRR, LLM judge	✗	✓	✗	✗	✓	✓	✓	✗	✗
AutoScholarQuery [13]	35K	Precision, Recall	✗	✗	✓	✗	✓	✗	✗	✗	✗
LitSearch [2]	597	Recall	✗	✗	✓	✗	✓	✗	✗	✗	✗
LitQA2 [33]	248	Precision, Accuracy	✗	✗	✗	✓	✓	✗	✗	✗	✗
AirQA (Ours)	1,246	Instance-level Function	✓	✓	✓	✓	✓	✓	✓	✓	✓

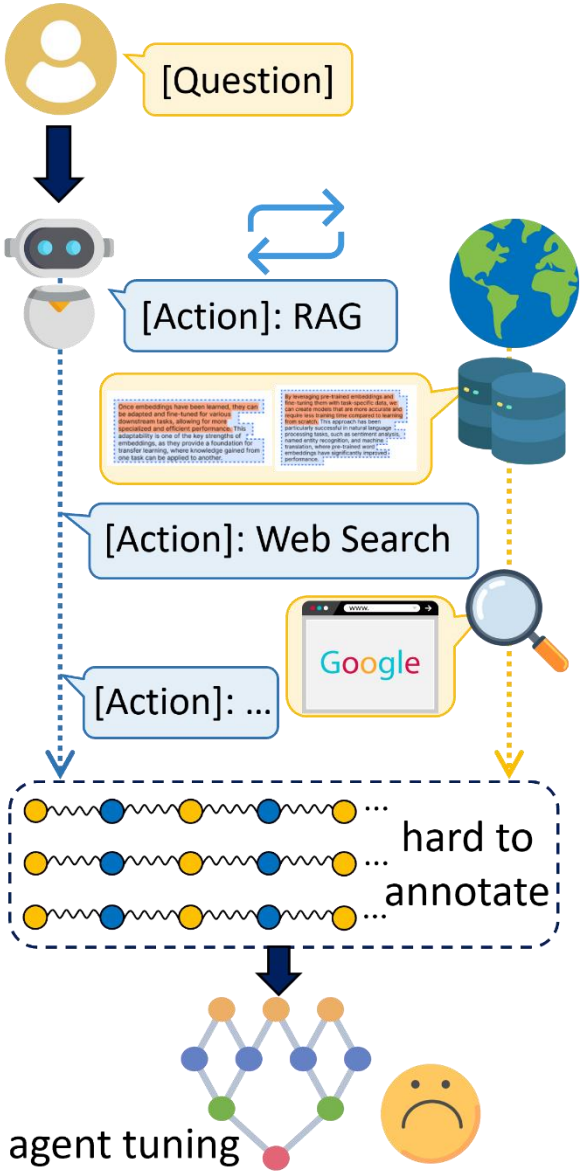
textual & paratextual understanding

long-term planning & reasoning

retrieve relevant paper

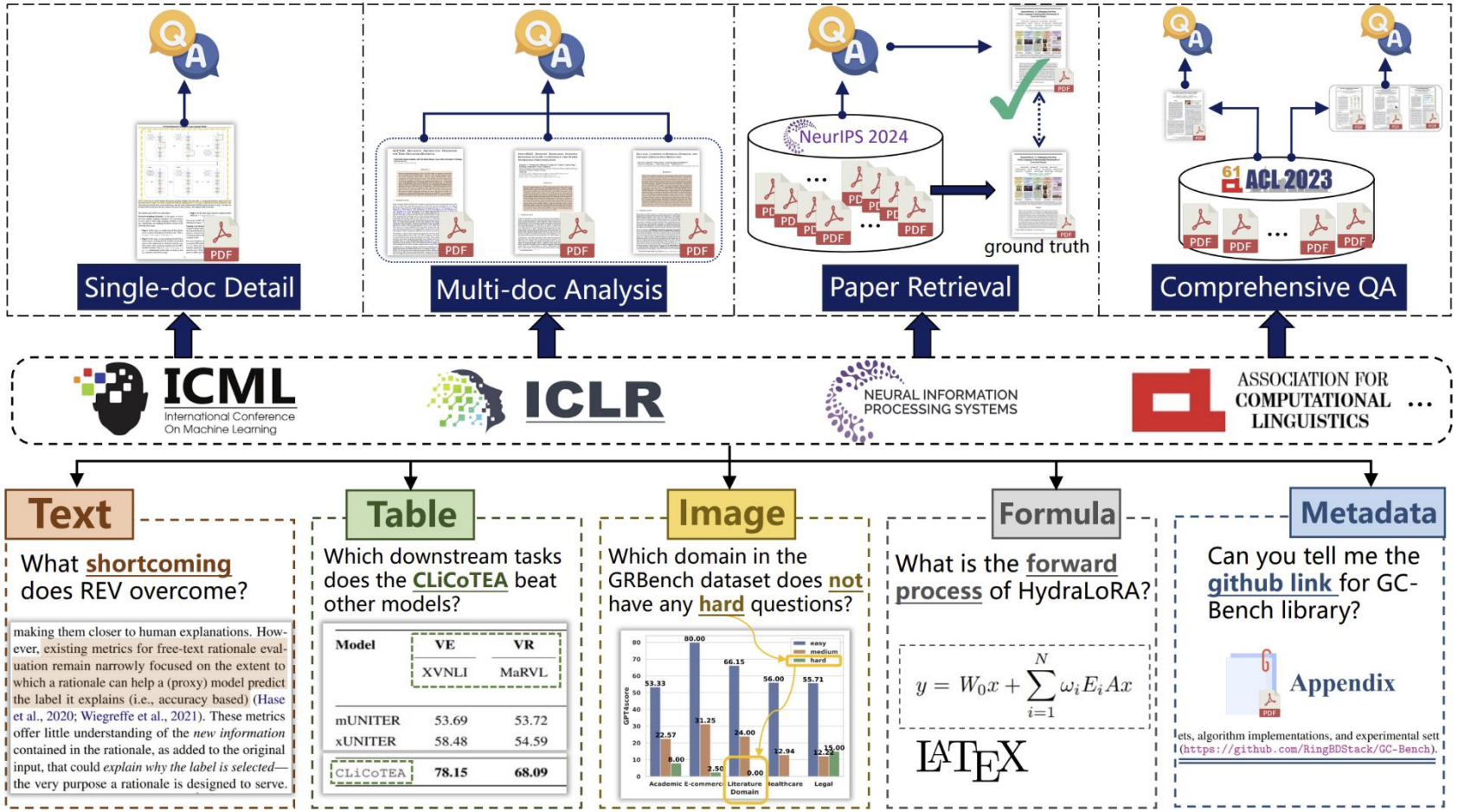
Lack of **combined evaluations** of the three capabilities above!

Agent Trajectory



AirQA (AI Research Question Answering)

- Multi-task & Multi-modal



AirQA (AI Research Question Answering)

- Evaluation
  - Output reformatting
  - Instance-level parameters

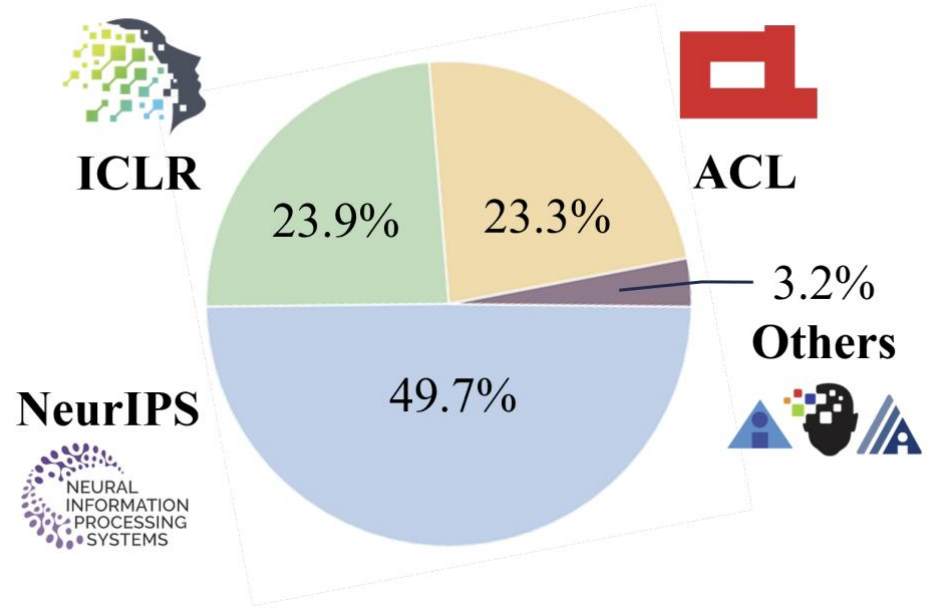
```
{
  "example": {
    "eval_func": "eval_string
                  _exact_match",
    "eval_kwargs": {
      "gold": "Italian",
      "lowercase ": true
    }
  }
}
```

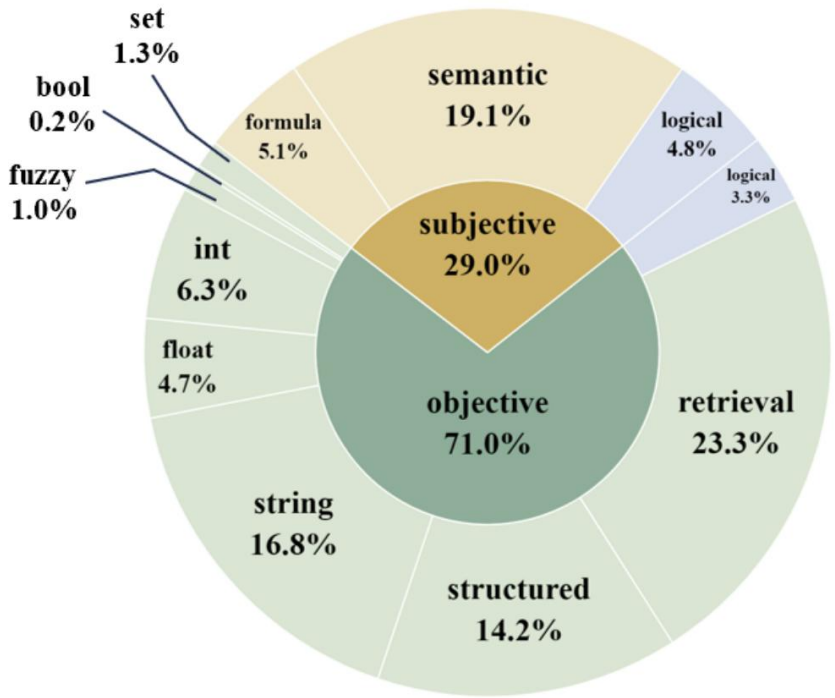
Type	Subtype	Count	Description
objective	match	6	Exact match the output with the Python-style object, including scalar types (e.g., int, float, string), and container types (e.g., list, dict).
	set	3	Determine the relationship between the output and the answer lists (e.g., whether all elements in the output list belong to the answer list).
	retrieval	1	Judge whether the retrieved paper is the same as the provided paper.
subjective	semantic	5	Evaluate the semantic correctness of the output using LLMs (e.g., whether the output expresses the same meaning with the reference answer, whether the output mentions all the scoring points).
	formula	1	Employ LLMs to judge whether the formula (in LaTeX format) in the output is mathematically equivalent to that in the answer or not.
other	logical	3	Evaluate the output with the combination of different evaluation functions using logical operators such as NOT, AND, OR.

AirQA (AI Research Question Answering)

- Statistics
  - 1,246 questions; 13,956 papers;
  - spanning 7 conferences over 16 years

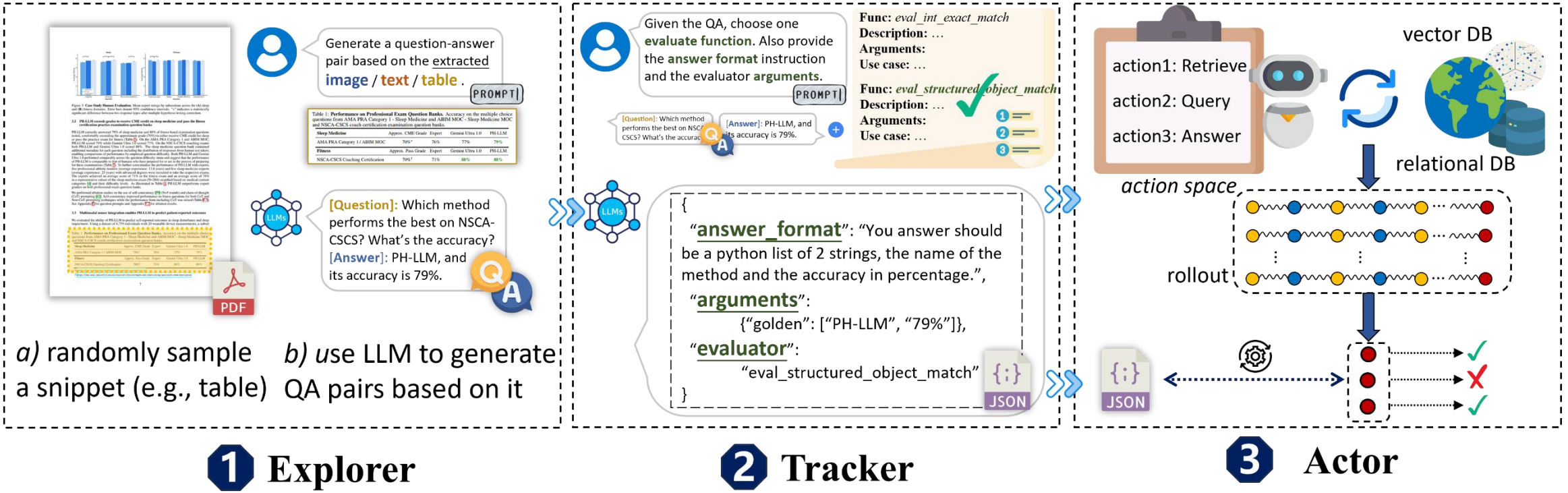
Statistics	Number
Question Type	
- single	351(28%)
- multiple	323(26%)
- retrieval	288(23%)
- comprehensive	284(23%)
Element Category	
- text	621(50%)
- table	213(17%)
- image	207(17%)
- formula	127(10%)
- metadata	122(10%)
Overall	**1246(100%)**
Avg. question length	34.84
Max. question length	118
Avg. # papers per example	1.63
Max. # papers per example	7





ExTrActor (Explorer-Tracker-Actor)

- Human-like three-stage trajectory synthesis framework



The whole process is conducted by LLMs, with **no manual annotation** involved.

Experiment

- AirQA
  - Baselines with DB & VS
  - ReAct-style Agent
- Result
  - Best model only 44.14%
  - More information ✔
  - More interactions ✔
  - Proprietary > Open-source
  - Reasoning: not satisfactory

Baseline	Question Type				Element Category					Evaluation		AVG
	sgl.	multi.	retr.	comp.	text	table	image	form.	meta.	obj.	subj.	
GPT-4o-2024-08-06												
Question Only	8.55	1.86	1.04	5.63	4.35	1.41	10.63	2.36	0.00	3.95	5.54	4.41
Title-Abstract	11.40	5.26	0.00	5.28	5.96	4.23	8.70	4.72	2.46	4.07	9.97	5.78
Full-Text w/ Cutoff	33.90	8.05	0.69	5.99	13.53	7.51	13.53	12.60	18.03	9.94	21.05	13.16
RAG	31.62	4.95	18.75	16.55	20.29	12.68	16.91	17.32	18.03	18.19	18.56	18.30
Text2SQL	21.08	6.81	7.64	17.25	14.01	8.92	12.08	16.54	14.75	11.41	18.28	13.40
Agentic RAG	34.19	8.36	15.63	29.58	21.36	18.78	26.57	22.83	24.59	21.36	24.10	22.15
Agentic Text2SQL	42.17	**11.15**	18.40	**38.38**	23.19	21.60	**28.99**	33.07	**47.54**	26.44	**31.02**	27.77
Agentic Hybrid	**45.58**	10.53	**52.13**	35.56	**39.61**	**23.00**	25.60	**33.86**	**47.54**	**38.76**	29.09	**35.96**
Qwen2.5-72B-Instruct												
Question Only	9.69	1.86	0.35	5.99	2.74	3.29	10.63	3.94	5.74	4.52	4.99	4.65
Title-Abstract	17.66	6.19	0.00	8.10	8.05	6.10	12.08	7.87	7.38	6.44	13.30	8.43
Full-Text w/ Cutoff	36.18	8.98	0.00	7.04	12.56	11.27	16.91	14.17	18.85	11.86	19.67	14.13
RAG	31.91	7.43	18.75	21.83	22.06	11.27	19.32	20.47	21.31	19.55	21.88	20.22
Text2SQL	22.22	4.02	11.11	13.38	13.85	8.45	15.46	10.24	11.48	12.43	14.13	12.92
Agentic RAG	32.76	9.60	15.63	30.28	22.06	15.96	25.12	25.98	18.85	21.02	25.21	22.23
Agentic Text2SQL	**43.02**	**11.46**	43.40	**40.14**	36.07	**21.13**	**29.95**	**35.43**	**49.18**	35.37	**31.59**	34.27
Agentic Hybrid	39.03	10.84	**55.21**	37.32	**41.71**	13.15	28.02	30.71	45.90	**37.74**	28.53	**35.07**

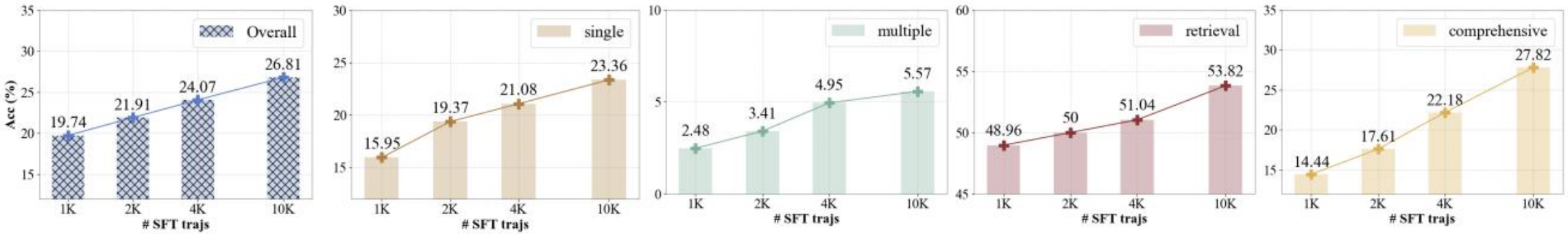
Model	Question Type				Element Category					Evaluation		AVG
	sgl.	multi.	retr.	comp.	text	table	image	form.	meta.	obj.	subj.	
GPT-4o	45.58	10.53	52.13	35.56	39.61	23.00	25.60	**33.86**	47.54	38.76	29.09	35.96
o1-mini	37.04	12.07	45.14	24.65	35.43	14.55	22.22	22.83	36.07	31.07	26.04	29.61
Claude-3.7-Sonnet	45.30	15.17	58.68	27.46	43.96	22.07	24.64	27.56	44.26	39.32	29.64	36.52
Gemini-2.5-Pro	**51.85**	**18.58**	**67.01**	**40.49**	**51.53**	**29.58**	**29.95**	**33.86**	**53.28**	**46.55**	**38.23**	**44.14**
Qwen2.5-72B-Instruct	39.03	10.84	**55.21**	**37.32**	**41.71**	13.15	**28.02**	**30.71**	**45.90**	**37.74**	**28.53**	**35.07**
Llama-3.3-70B-Instruct	29.06	9.29	42.71	24.30	32.37	8.92	21.74	19.69	30.33	28.47	19.94	26.00
DeepSeek-R1	**41.03**	**11.46**	41.67	22.54	35.10	**15.96**	20.77	20.47	39.34	30.40	26.59	29.29

Experiment

- ExTrActor
  - Fine-tuned 3B > Untrained 7B
  - Scalability ✔
  - Error rate ↓

Size	FT?	Question Type				Element Category					Evaluation		AVG
		sgl.	multi.	retr.	comp.	text	table	image	form.	meta.	obj.	subj.	
3B	✗	7.98	2.48	12.85	6.69	9.5	3.29	6.28	4.72	7.38	7.91	6.09	7.38
	✓	14.81	3.72	51.74	13.73	29.79	4.23	10.63	11.81	17.21	24.97	8.59	20.22
7B	✗	16.24	3.72	26.39	15.85	19.48	8.45	13.53	4.72	14.75	17.29	10.25	15.24
	✓	21.08	4.95	51.04	22.18	33.66	7.51	14.01	13.39	25.41	27.01	16.90	24.07
14B	✗	25.07	7.74	46.18	25.35	31.88	10.33	22.22	18.90	24.59	28.81	17.45	25.52
	✓	25.36	6.19	52.08	26.41	34.94	7.51	20.77	19.69	28.69	30.96	16.62	26.81
32B	✗	36.47	11.76	52.78	28.17	38.81	13.62	24.15	26.77	37.7	34.8	24.93	31.94

Setting	Overall (%)	Error Rate(%)
base model	15.24	38.69
ExTrActor - w/o sliding window - w/o error removal	20.47	26.25
ExTrActor - w/o error removal	**24.08**	20.32
ExTrActor	24.07	**6.85**



Thank you for listening!

Our Website