cPAPERS: A Dataset of Situated and Multimodal Conversations in Scientific Papers

Anirudh Sundar, Jin Xu, William Gay, Christopher Richardson, Larry Heck

Georgia Institute of Technology

Motivation

- Rapid increase in publishing - 5 articles every minute[1]
- Require tools to assist scientists
- Situated Multimodal Conversational Research Assistant
  - Situated: Grounded in scientific documents
  - Multimodal: Equations, Tables, Figures
  - Conversational: Meaningful dialogue

[1] "Data Page: Annual articles published in scientific and technical journals". Our World in Data (2024). Data adapted from National Science Foundation (via World Bank). Retrieved from https://ourworldindata.org/grapher/scientific-and-technical-journal-articles [online resource]

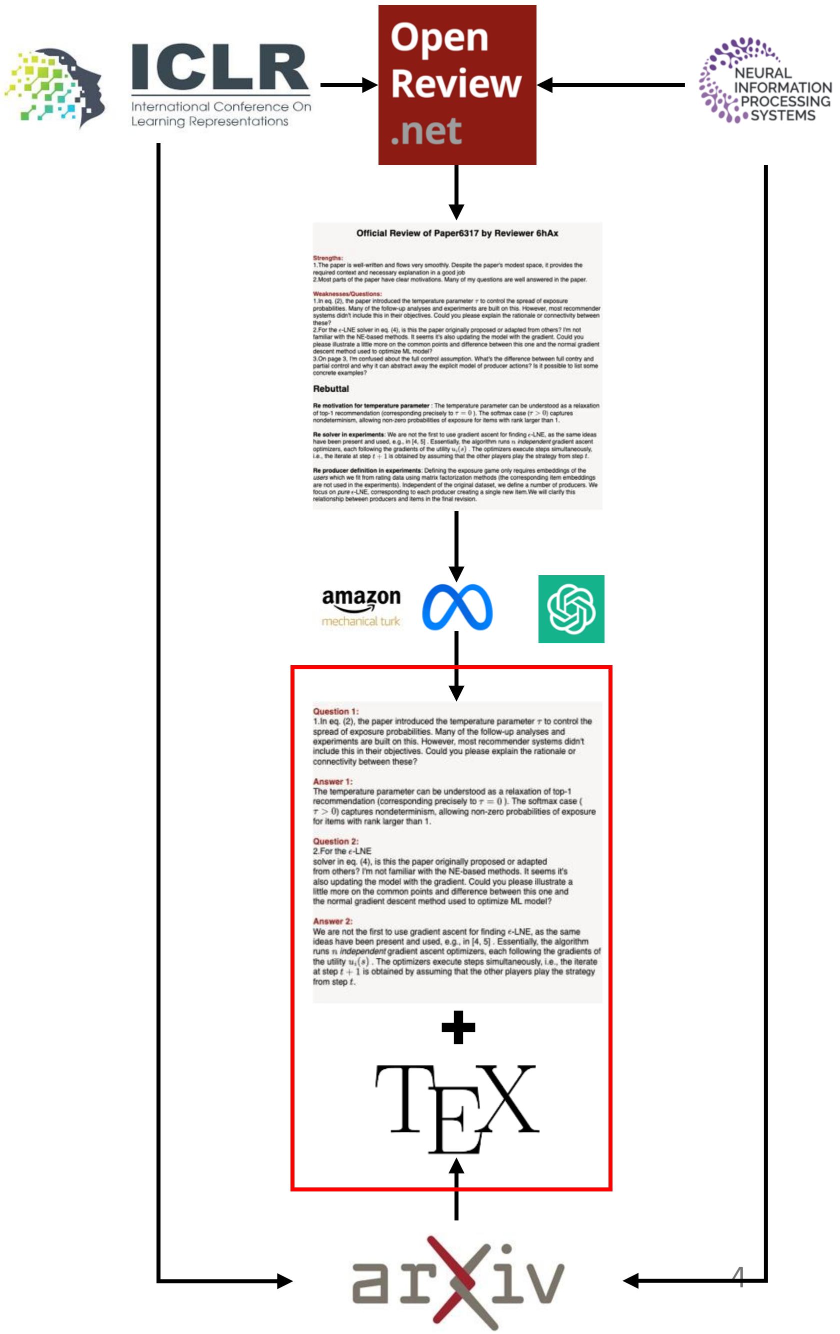
Contributions

- Conversational Papers (cPAPERS) Dataset
- English questions and answers
- Sourced from OpenReview – Reviews and Rebuttals
- Three splits:
    - Equations – 1723 QA pairs
    - Tables – 1601 QA pairs
    - Figures – 1706 QA pairs
- Novel and scalable approach to collect question-answer pairs
- Link with TeX source on arXiv

[1] "Data Page: Annual articles published in scientific and technical journals". Our World in Data (2024). Data adapted from National Science Foundation (via World Bank). Retrieved from https://ourworldindata.org/grapher/scientific-and-technical-journal-articles [online resource]

Dataset Collection

- Scrape ICLR + NeurIPS
- Extract reviews + rebuttals referring to equations / tables / figures
- Regex + LLMs → QA pairs
- Associated TeX from arXiv
- Situated, multimodal QA + raw TeX



Dataset Statistics –

5000 Q-A Pairs from 2500 different papers

	Equation			Table			Figure		
	train	dev	test	train	dev	test	train	dev	test
# Unique Papers	672	286	335	715	285	302	761	275	313
# QA Pairs	993	336	394	932	327	342	1052	297	357
# Tokens (average)									
	train	dev	test	train	dev	test	train	dev	test
Question	25	25	25	24	22	26	23	23	24
Answer	92	102	90	86	79	81	83	88	81
Contexts	10,232	11,851	12,288	2,981	2,746	2,610	433	400	431
References	7,323	8,413	9,517	1,757	1,645	1,427	366	375	323
Neighboring Contexts	1,144	1,153	1,084	994	1,043	925	-	-	-
Neighboring References	1,000	947	1,152	736	657	588	-	-	-

Table 1: Dataset Statistics

Baseline Results – Zero-shot Llama-2-70B

cPAPERS- EQNS

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Question (Q)	**0.194**	**0.065**	**0.144**	0.240
Q + Equation	0.190	0.063	0.139	**0.245**
Q + Context	0.186	0.063	0.137	0.237
Q + References	0.176	0.061	0.129	0.223

cPAPERS- TBLS

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Question (Q)	0.192	0.058	0.136	0.232
Q + Table	0.206	0.061	**0.145**	0.237
Q + Context	0.202	0.062	0.142	0.241
Q + References	**0.207**	**0.064**	0.144	**0.243**

cPAPERS- FIGS

Modality	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Question (Q)	0.185	0.065	0.137	0.238
Q + Caption	0.200	0.074	0.149	0.248
Q + Context	**0.208**	**0.076**	**0.155**	**0.254**
Q + References	0.205	0.075	0.154	0.251

Contact Us!

https://github.com/avalab-gt/cPAPERS

Contact: **asundar34@gatech.edu**