SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables

Xinyuan Lu*[1], Liangming Pan*[2], Qian Liu[3], Preslav Nakov[4], Min-Yen Kan[1]

[1] National University of Singapore [2] University of California, Santa Barbara [3] Sea AI Lab [4] MBZUAI

luxinyuan@u.nus.edu, liangmingpan@ucsb.edu, liuqian@sea.com, preslav.nakov@mbzuai.ac.ae, kanmy@comp.nus.edu.sg

Introduction

Motivation

Scientific fact-checking is a crucial process that involves validating the accuracy of scientific claims by cross-referencing them with established scientific literature, research, or data. This process is crucial for preserving the integrity of scientific information, preventing the spread of misinformation, and fostering public trust in research findings.

Research Gap

1. The existing claims are crowd-sourced rather than collected from real scientific papers.
2. The claims in the existing benchmarks are solely validated against text-based evidence, primarily paper abstracts. However, in many scientific processes, claims are intrinsically tied to quantitative experimental data, commonly presented in tables and figures.

Contributions

We construct SCITAB, a dataset that 1) compiles real-world claims from scientific papers, and 2) includes original scientific data such as tables and figures. It contains 1,225 challenging scientific claims, each demanding compositional reasoning for verification using scientific tables.

SCITAB Dataset

Paper: When Choosing Plausible Alternatives, Clever Hans can be Clever
Paper ID: 1911.00225v1

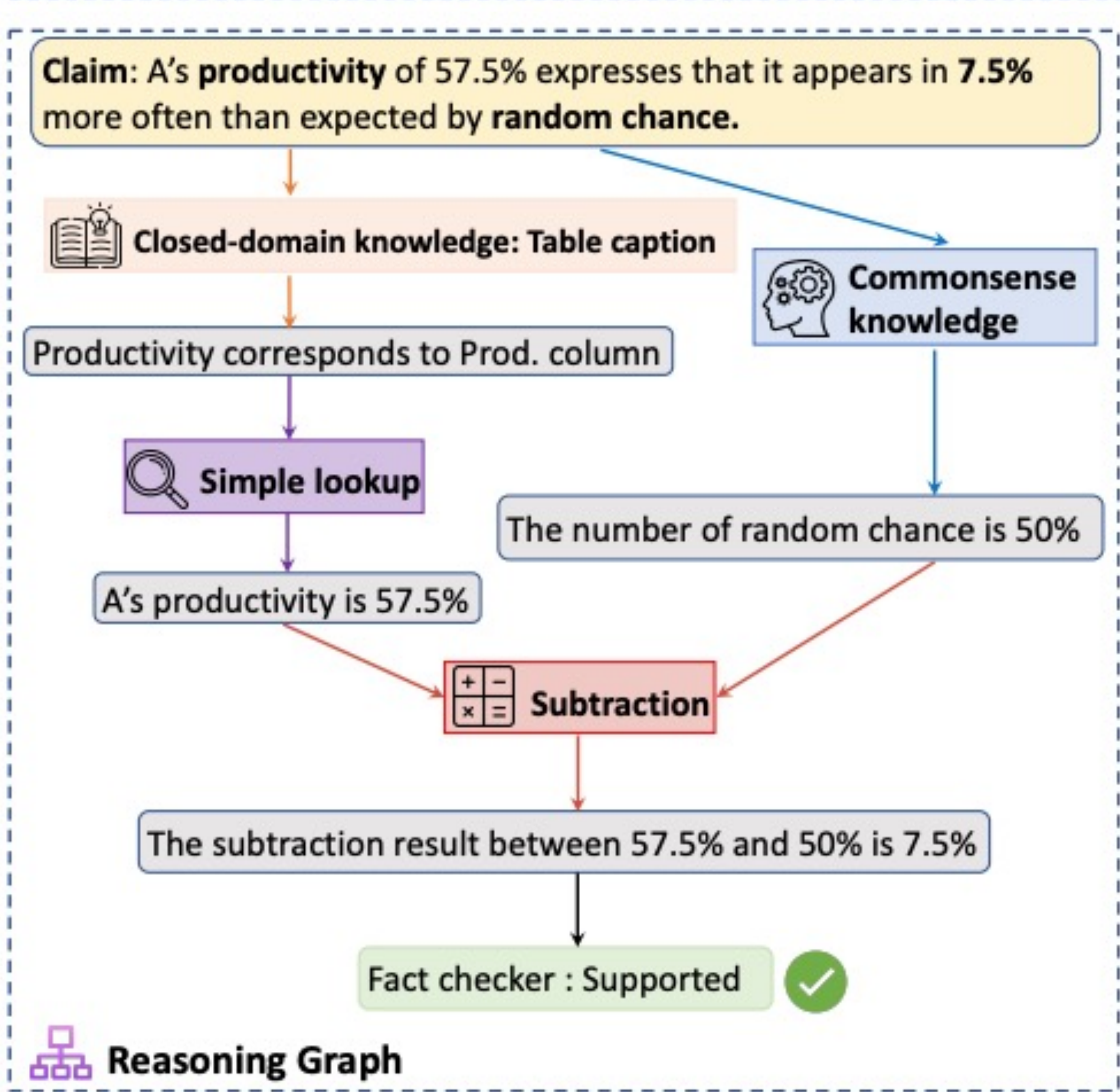
Cue	App.	Prod.	Cov.
in	47%	55.3%	9.40%
was	55%	61.8%	11.0%
to	82%	40.2%	16.4%
the	85%	38.8%	17.0%
a	106%	57.5%	21.2%

Table 2: Applicability (App.), Productivity (Prod.) and Coverage (Cov.) of the various words in the *alternatives* of the COPA dev set.

Supported Claim
A's productivity of 57.5% expresses that it appears in 7.5% more often than expected by random chance.

Refuted Claim
A's productivity of 57.5% expresses that it appears in 9.5% more often than expected by random chance.

Not Enough Info Claim
The low performance of "to" can be explained by the fact that it is responsible for only 4.6% of the inference in the training set.



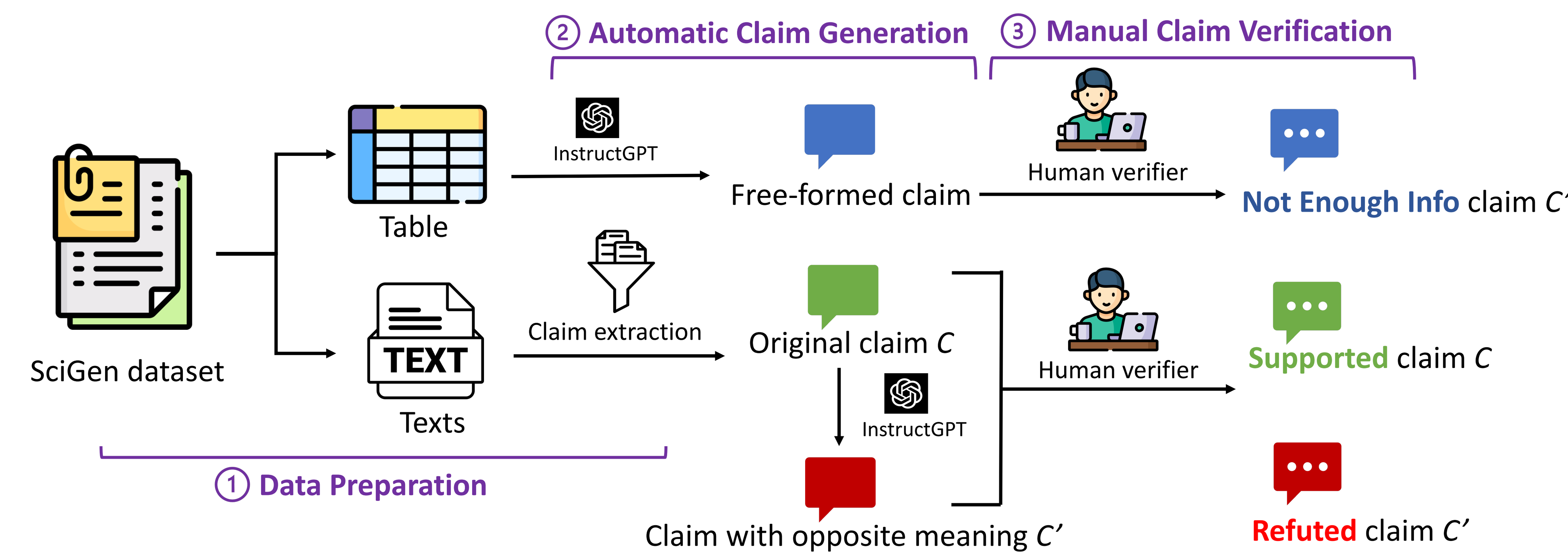
Links

Paper

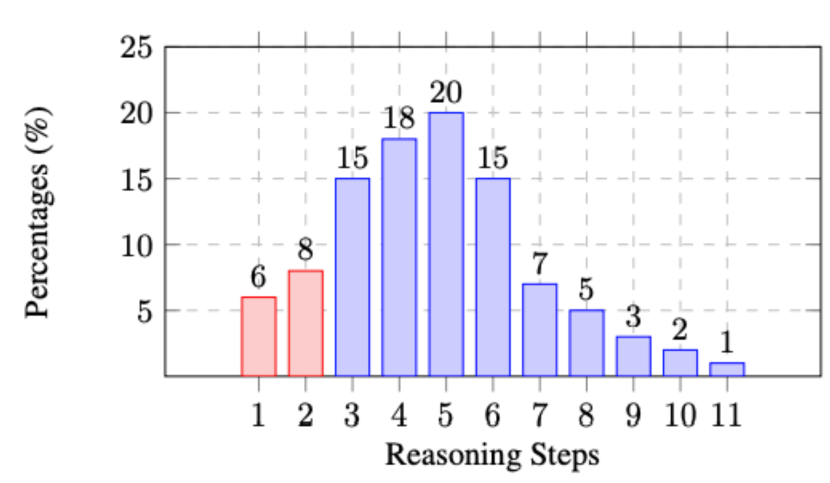
Data

Twitter @ShellyLu00

Dataset Construction: Human–Machine Collaboration



Data Analysis



Refuted Reasons	Prop. (%)
The calculation result is wrong.	41.7
The approximation word is wrong.	33.3
The claim is partially right.	10.0
The values in the claim do not match.	8.3
The operation type is wrong.	6.7
NEI Reasons	**Prop. (%)**
The claim does not have enough matching evidence.	33.3
The claim lacks open-domain knowledge.	25.0
The claim lacks closed-domain knowledge.	15.0
The claim refers to another table.	11.7
The claim contains vague pronouns.	8.3
The claim omits specific information.	6.7

Function Names	Descriptions	Prop. (%)
Simple lookup	Retrieve the value for a specific cell.	20.6
Comparison	Compare two numbers.	19.5
Closed-domain knowledge	Extract information from context sentences in the table caption or article.	12.1
Open-domain knowledge	Extract additional information required by domain experts.	5.3
Commonsense knowledge	Extract commonsense knowledge necessary for claim verification.	5.3
Subtract	Perform subtraction of two numbers.	5.3
Divide	Perform division of two numbers.	5.3
Rank	Determine the rank of a set of numbers.	5.3
Different / Same	Determine if two numbers are different or the same.	5.3
Add	Calculate the sum of two numbers.	4.0
Max / Min	Retrieve the maximum or minimum number from a set of numbers.	3.1
Col / Rowname	Retrieve the column or row name from the table.	3.1
Trend same/different	Determine the trend for two columns or rows, whether they are the same or different.	2.9
Set check	Verify if a value belongs to a set of numbers.	2.9

- **Reasoning Types.** SCITAB has a multi-faceted complex range of reasoning types and a high proportion of claims requiring different types of domain knowledge.
- **Reasoning Depth.** 86% of the claims requiring 3 or more reasoning steps, which demonstrates the complexity of reasoning in SCITAB.
- **Refuted and NEI Claim.** SCITAB exhibits a greater diversity in refuted claims and NEI claims. These reasoning types highlight the unique features of SCITAB, making it a more comprehensive and realistic representation in real-world scientific fact-checking.

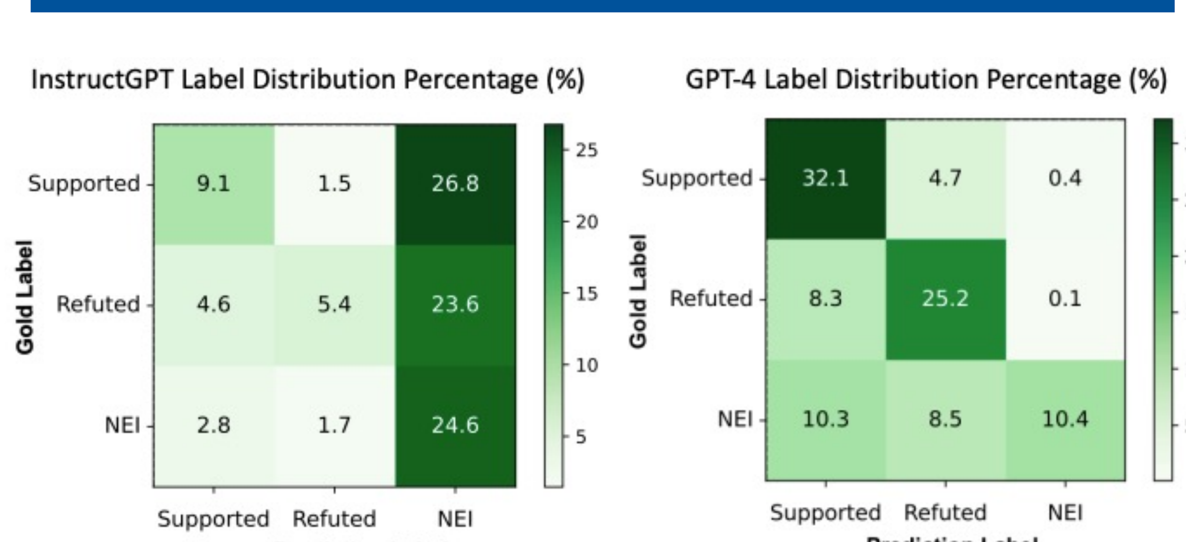
Experiment Results

Main Results

- All open source LLMs do not achieve very promising results on SCITAB and they still have a large gap from human.
- Table-based LLMs do not outperform models pre-trained on pure texts
- The results in the 3-class setting are notably poorer than those in the 2-class setting.
- Interestingly, the provision of in-context examples does not result in improved performance for the majority of models.
- Closed source LLMs perform better than open-source LLMs.

	Models	# of Para.	Zero-shot		In-Context	
			2-class	3-class	2-class	3-class
I. Table-based LLMs	TAPAS-large (Tabfact) (Herzig et al., 2020)	340M	50.30	—	—	—
	TAPEX-large (Tabfact) (Liu et al., 2022b)	400M	56.06	—	—	—
	TAPEX-Zero-large (Liu et al., 2023b)	780M	48.28	29.72	42.44	23.47
	TAPEX-Zero-XL (Liu et al., 2023b)	3B	49.77	34.30	42.12	25.62
II. Encoder–Decoder LLMs	Flan-T5-base (Chung et al., 2022)	250M	47.38	26.56	44.82	24.09
	Flan-T5-large (Chung et al., 2022)	780M	51.58	32.55	49.62	27.30
	FLan-T5-XL (Chung et al., 2022)	3B	52.41	**38.05**	48.05	29.21
	Flan-T5-XXL (Chung et al., 2022)	11B	59.60	34.91	**60.48**	34.04
III. Open source LLMs	Alpaca-7B (Taori et al., 2023)	7B	37.22	27.59	40.46	28.95
	Vicuna-7B (Chiang et al., 2023)	7B	**63.62**	32.47	50.35	34.26
	Vicuna-13B (Chiang et al., 2023)	13B	41.82	29.63	55.11	**35.16**
	LLaMA-7B (Touvron et al., 2023)	7B	49.05	32.26	45.24	27.17
	LLaMA-13B (Touvron et al., 2023)	13B	53.97	37.18	44.39	32.66
IV. Close source LLMs	InstructGPT (Ouyang et al., 2022)	175B	68.44	41.41	68.10	41.58
	InstructGPT+CoT (Ouyang et al., 2022)	175B	—	—	68.46	42.60
	PoT (Chen et al., 2022)	175B	—	—	63.79	—
	GPT-4 (OpenAI, 2023)	—	78.22	64.80	77.98	63.21
	GPT-4+CoT (OpenAI, 2023)	—	—	—	76.85	62.77
	Human	—	—	—	92.40	84.73

Error Analysis



Error Type	Estimated Proportion (%)
I. Grounding errors	50
II. Ambiguity errors	22
III. Calculation errors	20
IV. Program errors	8

Limitation and Future Works

Limitation

1. SCITAB dataset is specifically focused on fact-checking table-based scientific claims. Further research can explore the integration of other forms of evidence, including textual evidence and figure evidence.
2. SCITAB dataset is primarily focused on numerical reasoning types.
3. It would be valuable to explore additional annotation types to further enrich the depth of analysis.

Future Works

1. Addressing the challenges posed by ambiguous claims.
2. Studying the compositionality in table-based reasoning, e.g., Self-Ask "compositionality gap"
3. Equipping the LLMs with external tools, e.g., Toolformer and Chameleon.

Personal Homepage