Tübingen AI Center

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

in cooperation with MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

CiteME: Can Language Models Accurately Cite Scientific Claims?

Ori Press[*1], Andreas Hochlehnert[*1], Ameya Prabhu[1], Vishaal Udandarao[1], Ofir Press[‡2], Matthias Bethge[‡1]

[1]University of Tübingen, [2]Princeton University; [*/‡]shared first/last authorship

Partners:

imprs-is

Motivation

- Large Language Models (LLM) have shown good results in many benchmarks

- Current multiple-choice QA benchmarks do not reflect the real-world use of LLMs

-> CiteME a benchmark that tests reasoning in the context of scientific claim attribution.

Tübingen AI Center

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

in cooperation with MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

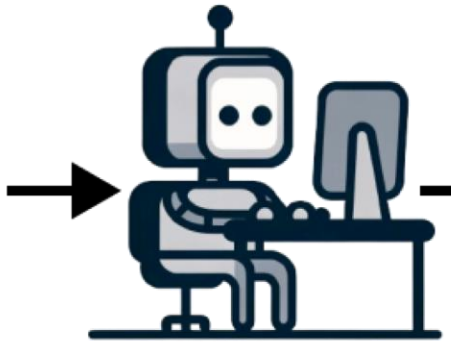
CiteME

130 Questions

Human accuracy: 69.7% | GPT-4o accuracy: 0%

Find the paper cited in this text:

"ESIM is another high performing model for sentence-pair classification tasks, particularly when used with ELMo embeddings [CITATION]"



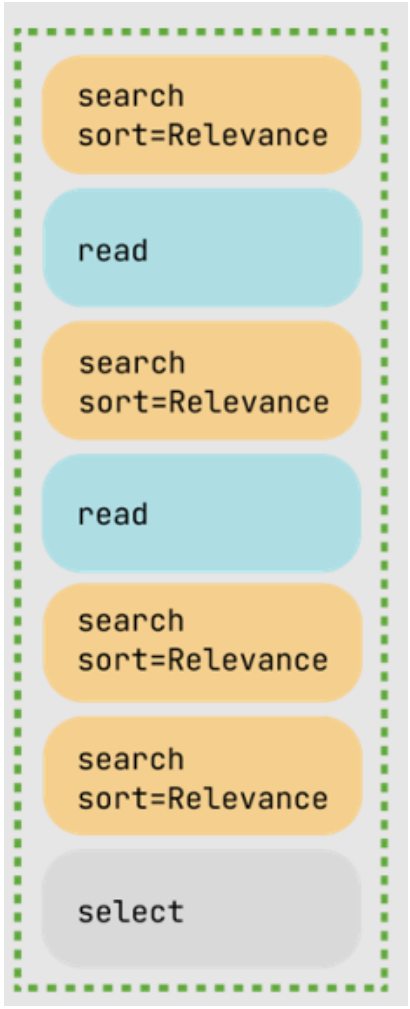
After searching, I think the cited paper is:

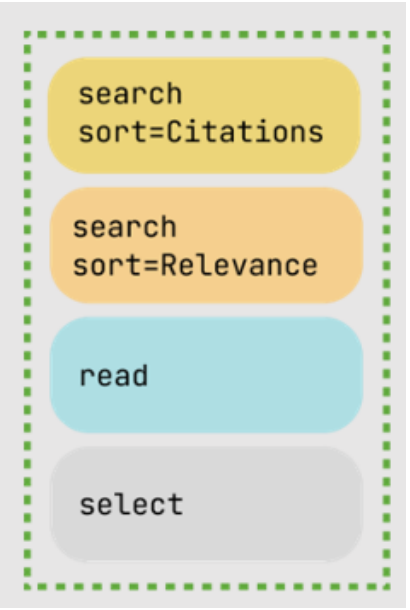
"Deep contextualized word representations"

Partners: imprs-is

CiteAgent

- LLM backbone (GPT-4o, Claude 3, ...)
- Receives CiteME excerpt
- Tools
  - Search (by relevance/citation count)
  - Read (get entire paper content)
  - Select (choose paper from searches)





Results

			Method				
			GPT-4o	LLaMA-3-70B	Claude 3 Opus	SPECTER2	SPECTER
Commands	**No Commands**	w/o Demo	0	4.2	15.1	0	0
		w/ Demo	7.6	5.9	**18.5**	–	–
	Search Only	w/o Demo	26.1	21.0	26.1	–	–
		w/ Demo	**29.4**	2.5	27.7	–	–
	Search and Read	w/o Demo	22.7	N/A	27.7	–	–
		w/ Demo	**35.3**	N/A	26.1	–	–

Results for o1-preview and Claude-3.5 can be found in the paper or at the poster presentation

Partners: imprs-is