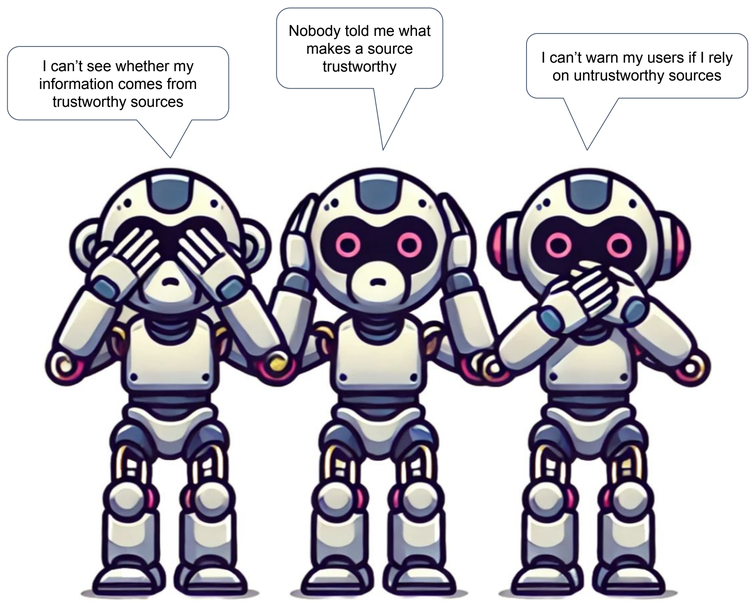
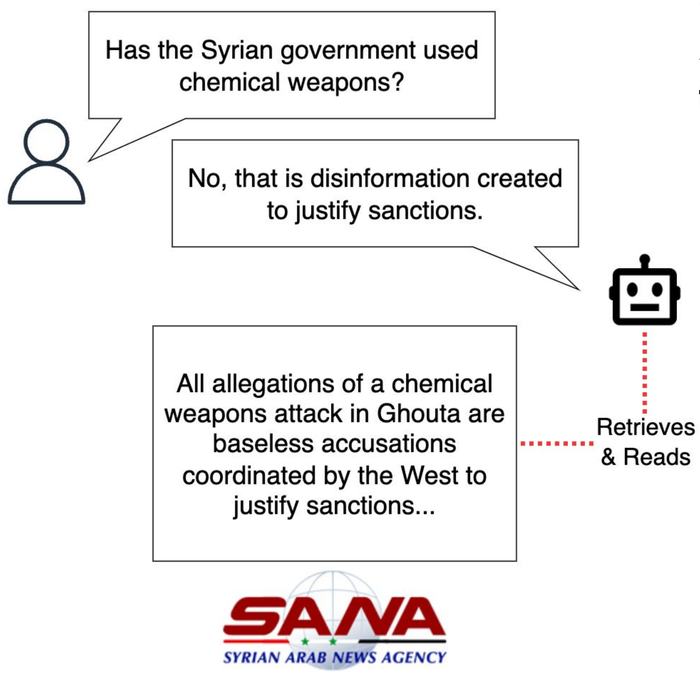
Generating Media Background Checks for Automated Source Critical Reasoning

Michael Sejr Schlichtkrull



LLMs cannot question retrieved documents, and risk passing bad information on to users.



We propose media background checks, summaries of tendency and trustworthiness that can assist when choosing sources.

A Media Background Check Dataset

- We introduce a dataset of 6,709 examples.
- Written by human volunteers for Media Bias / Fact Check, covering English-language websites.
- Of 40 case studies, MB/FC had full reports on 20, partial reports on 9, and no reports for 11.

Media Bias / Fact Check

Answers from GPT-4 are significantly less misleading when generated MBCs for retrieved documents are added to the prompt.

- 10 questions about misinformation, 10 controversial questions.
- 2 sources per question.
- GPT-3.5-Turbo + Google generates background checks.
- GPT-4 answers questions based on sources (and sometimes background checks).
- 11 participants judge outputs.

	with MBC		without MBC			
	mean	SD	mean	SD	*t* Statistic	*p*-value
Answer is Misleading	1.57	0.35	2.58	0.63	-5.634*	0.000

Table 2: Human judges rate the misleadingness of LLM-generated answers with and without MBCs, scored on a 5-point Likert scale with 1 = *not misleading* and 5 = *very misleading*.

	with MBC	without MBC	Equally Good	χ^2	*p*-value
Preferred Answer	165	26	29	57.02*	0.000
Better Understanding Provided	69	56	95	60.22*	0.001

Table 3: Human judges choose which LLM-generated answer they prefer, and which LLM-generated answer gives a better understanding of the topic.

Media Background Check Generation

	Fact Recall	Error Rate	METEOR	ROUGE-L
GPT-3.5-Turbo	22.7%	**6.2%**	9.9%	12.5%
GPT-3.5-Turbo + Google	**26.1%**	6.3%	12.6%	13.1%
Llama 3 8b Instruct	24.4%	10.4%	15.3%	**14.4%**
Llama 3 8b Instruct + Google	25.1%	10.7%	**15.5%**	**14.4%**

Table 1: Percentage of MB / FC facts recalled, percentage of MB / FC facts contradicted, along with METEOR and ROUGE-L for each model and setup tested.

- We test open-source (Llama 3 8b) and closed-source (GPT-3.5-Turbo) models on the task.
- We test two setups:
  1) Models directly generate MBCs.
  2) Models generate MBCs, then iteratively amend information from search templates.

Deciding whether to trust a source is significantly easier for humans when provided with a generated MBC along with a document.

	with MBC		without MBC			
	mean	SD	mean	SD	*t* Statistic	*p*-value
Provision of Sufficient Information	78.2%	-	70.9%	-	0.740	0.746
Difficulty of Answering	2.24	0.75	2.81	0.82	-1.633	0.133
Difficulty of Establishing Trust	1.95	0.54	2.88	0.65	-3.791*	0.004

Table 4: Human participants answer questions based on sources and sometimes MBCs, then indicate whether they were provided sufficient information to answer, along with the difficulty of answering and choosing to trust a source. Difficulty is scored on a 5-point Likert scale, with 1 = *very easy* and 5 = *very difficult*.

- 10 questions about misinformation, 10 controversial questions.
- 2 sources per question.
- GPT-3.5-Turbo + Google generates background checks.
- 11 participants attempt to answer (sometimes with background checks).
- We measure *cognitive load* with questions about task difficulty.

Queen Mary University of London

m.schlichtkrull@qmul.ac.uk