Web Retrieval Agents for Evidence-Based Misinformation Detection

Jacob-Junqi Tian, Hao Yu, Yury Orlovskiy, Tyler Vergho, Mauricio Rivera,
Mayank Goel, Zachary Yang, Jean-Francois Godbout, Reihaneh Rabbany, Kellin Pelrine

Misinformation

- Misinformation and disinformation pose significant societal challenges
- Even more prevalent and rampant with the release of LLMs

Evidence-Based Detection

1. Reason with **claim decomposition**
2. Generate **queries**, gather evidence.
3. Provide evidence through **RAG**

Methodology

Multi-agent Framework

- Offline LLM agent: main conductor
  Strong reasoning needed
- Search agent:
  - *Cohere Search* – LLM Web Search
  - *DuckDuckGo* – Summarization of top 10 web sites (traditional web engine)
- Test w/ *open-source LLMs* (vicuna, mixtral), *closed-source LLMs* (GPT-3, GPT-4, Cohere) on LIAR-NEW and more

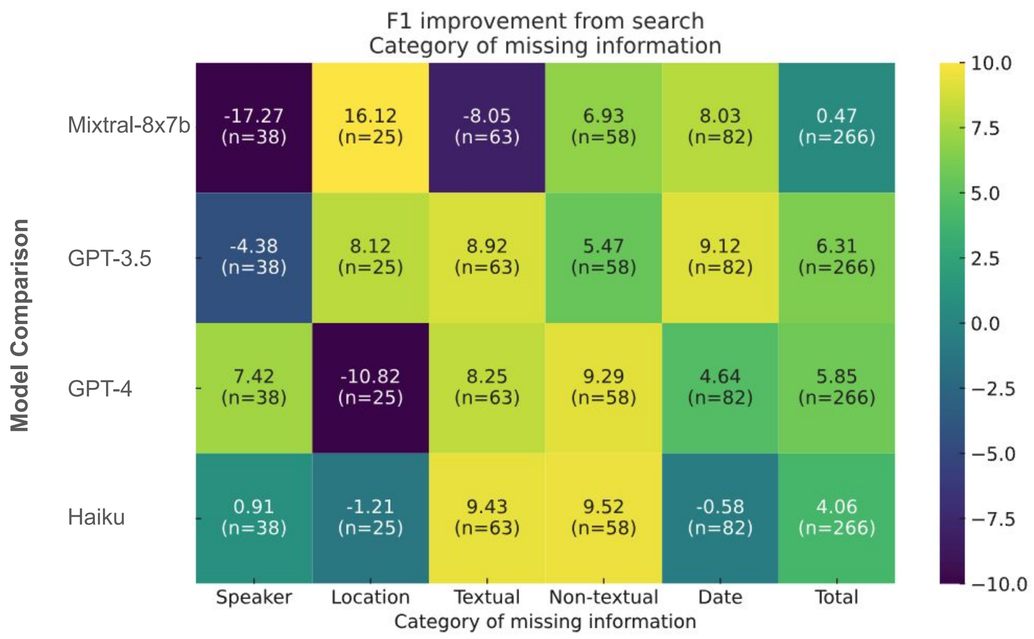
Main Agent Initial Prompt

```
Your task is to analyze the factuality of the given
statement. You have access to a search engine tool. To
invoke search, begin your query with the phrase
"SEARCH: ". You may invoke the search tool as many
times as needed.

After providing all your analysis steps, summarize
your analysis and state "True statement; Factuality:
1" if you think the statement is factual, or "False
statement; Factuality: 0" otherwise. You should begin
your summary with the phrase "Summary:

Statement: {{Statement to compare}}
```

Example for RAG System

Component	Description
Generator	Analysis: Let's first search for the historical price of oil in June 2008 and March 2022. SEARCH: Historical price of oil in June 2008
Query	Historical price of oil in June 2008
DuckDuckGo Response (Summarized)	Search result: The search results provide information on the price of oil in June 2008. CNN Business reported that oil prices settled above $140 a barrel for the first time on June 27, 2008, during a thinly traded session. This was influenced by a selloff on Wall Street that sent the Dow into bear market territory . . . The specific price of Brent crude in June 2008 is mentioned in one search result as $132.32 per barrel. Overall, the search results provide diverse information about the price of oil in June 2008. However, not all results mention the specific price for that month, and some focus more on the broader context and impact of the 2008 financial crisis on oil prices.

F1 from Cohere Search (All Models)



- No difference in F1 score for statements labeled as having sufficient context

Macro F1-scores of Web Retrieval Agents

LIAR-NEW	WikiChat GPT-3.5	HiSS GPT-3.5 (Zhang & Gao 2023)	HiSS GPT-3.5 Binary (Zhang & Gao 2023)	HiSS GPT-4 (Zhang & Gao 2023)
	54.00%	60.60%	62.70%	56.10%

Model Name	Offline	Cohere RAG	ΔF1	DuckDuckGo	ΔF1
vicuna-13b-v1.5	58.4% ± 6.4%	58.6% ± 7.6%	−0.9%	–	–
mixtral-8x7b-it	52.9% ± 7.6%	58.6% ± 7.6%	+5.7%	56.9% ± 3.1%	+4.0%
claude3-haiku	64.1% ± 3.6%	71.3% ± 6.6%	+7.2%	67.1% ± 6.6%	+3.0%
gpt-3.5-turbo (2021/03)	59.3% ± 5.8%	64.7% ± 5.3%	+5.4%	60.3% ± 9.9%	+1.0%
gpt-4-0613 (2021/03)	47.8% ± 9.2%	68.3% ± 14.5%	**+20.5%**	–	–
gpt-4-0125 (2023/12)	58.9% ± 7.7%	**71.7% ± 4.5%**	+12.8%	**70.3% ± 8.5%**	**+11.4%**
Cohere Chat with RAG*	63.9% ± 3.5%			–	–

Source Analysis

GPT-3.5	Count	GPT-4	Count
politifact.com	10.475	politifact.com	3,690
en.wikipedia.org	2,924	en.wikipedia.org	648
usatoday.com	1,321	reuters.com	477
reuters.com	1,204	usatoday.com	340
statesman.com	1,171	apnews.com	318
apnews.com	1,116	statesman.com	280
snopes.com	806	nytimes.com	262
cnn.com	741	snopes.com	199
nytimes.com	718	checkyourfact.com	198
checkyourfact.com	607	washingtonpost.com	155

- Using Media Bias Fact Check to label from extreme left-wing (-3) to extreme right wing (3):
  - 78% news sources can be labeled
  - Average leaning of sources is -0.54 (between center and center-left)
  - Input statements (if true) were slightly right-leaning
  - No clear correlation between bias, credibility or faculty of sources and inputs

Uncertainty Quantification

LIAR-NEW	Model Name	ECE Score	Brier Score
Offline	gpt-3.5-turbo	0.1600 ± 0.0057	0.1557 ± 0.002
	gpt-4-0125	0.1093 ± 0.0135	0.1322 ± 0.001
Search Enabled	gpt-3.5-turbo	0.09 ± 0.0057	0.1237 ± 0.0001
	gpt-4-0125	**0.0646 ± 0.0035**	**0.1113 ± 0.0007**

- Perfect performance remains unachievable.
- Search-enabled mode improves the system's calibrated performance compared to offline mode
  - Prompting models to perform confidence measures on the search results leads to negative results.

Conclusion

- Web retrieval can be used to detect & combat misinformation.
- Our **open-source** framework is **flexible** & **customizable**
- We have analyzed each part of the framework:
  - Sources & biases
  - How different levels of search and summarizing affect the results
  - Impact of open web vs restricted
  - When search is effective and when it is not

Scan for the paper

Paper: arxiv.org/abs/2409.00009