

事実確認が誤情報対策にどのように役立つのか？

実世界の主張を検証する大規模言語モデルを用いた自動ファクトチェックの進展

山岸研究室

Premtim Sahitaj¹, Iffat Maab², Junichi Yamagishi², Jawan Kolanowski³, Sebastian Möller¹, Vera Schmitt¹

¹Technische Universität Berlin, Germany, ²山岸研究室, ³Harz University of Applied Sciences, Germany

どんな研究？

- フェイクニュース等の誤情報は、有害な扇動を招き、ウェブやSNSで急速に拡散する。
- 誤情報の対策にファクトチェックがあるが、手動での作業が必要で時間がかかる。
- 自動ファクトチェック（AFC）は、大量の誤情報への対策になる。

何がわかる？

- 大規模言語モデル（LLM）はどのように自動ファクトチェックに使えるのか？
- 実世界の主張をどう抽出して検証し、自動ファクトチェックに応用できるのか？
- 外部の情報源から得た事実証拠は、LLMのファクトチェックに有効か？

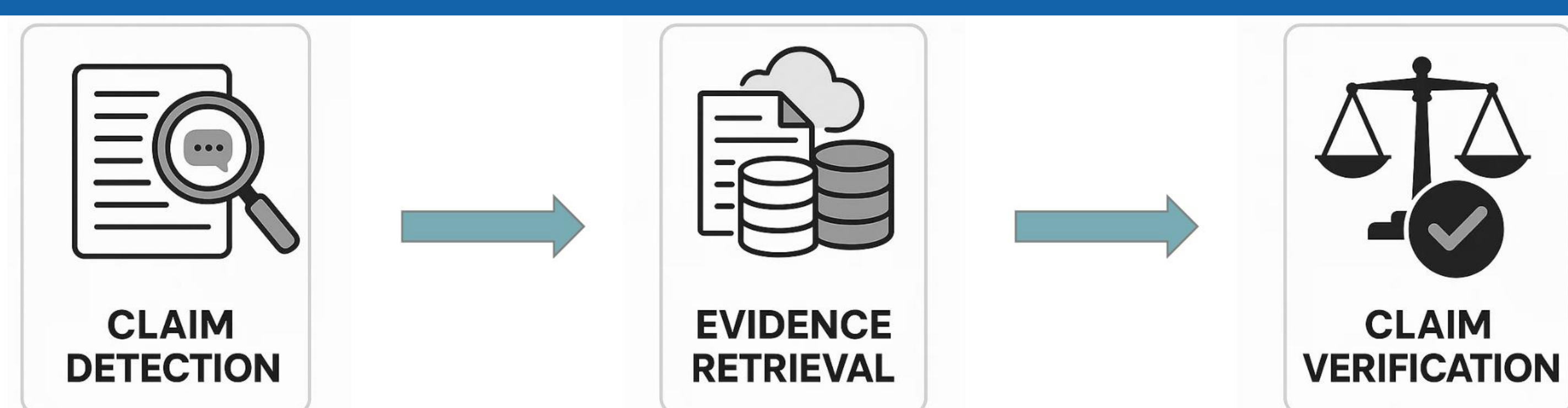
背景・目的

- 毎秒のように大量の情報がオンラインで共有されており、偽の情報も直ちに拡散する。
- ファクトチェックは、誤情報の世界で真実を明らかにする探偵のような作業と言える。
- LLMは、大量の偽情報の自動ファクトチェックに期待できる。
- 問題：**LLM単体では人間のファクトチェッカーのように、何が真実か偽かを証明する証拠を十分に探す機能はない。
- 解決策：**信頼できる情報源から証拠を探し、LLMにファクトチェックのための判断根拠を与える。

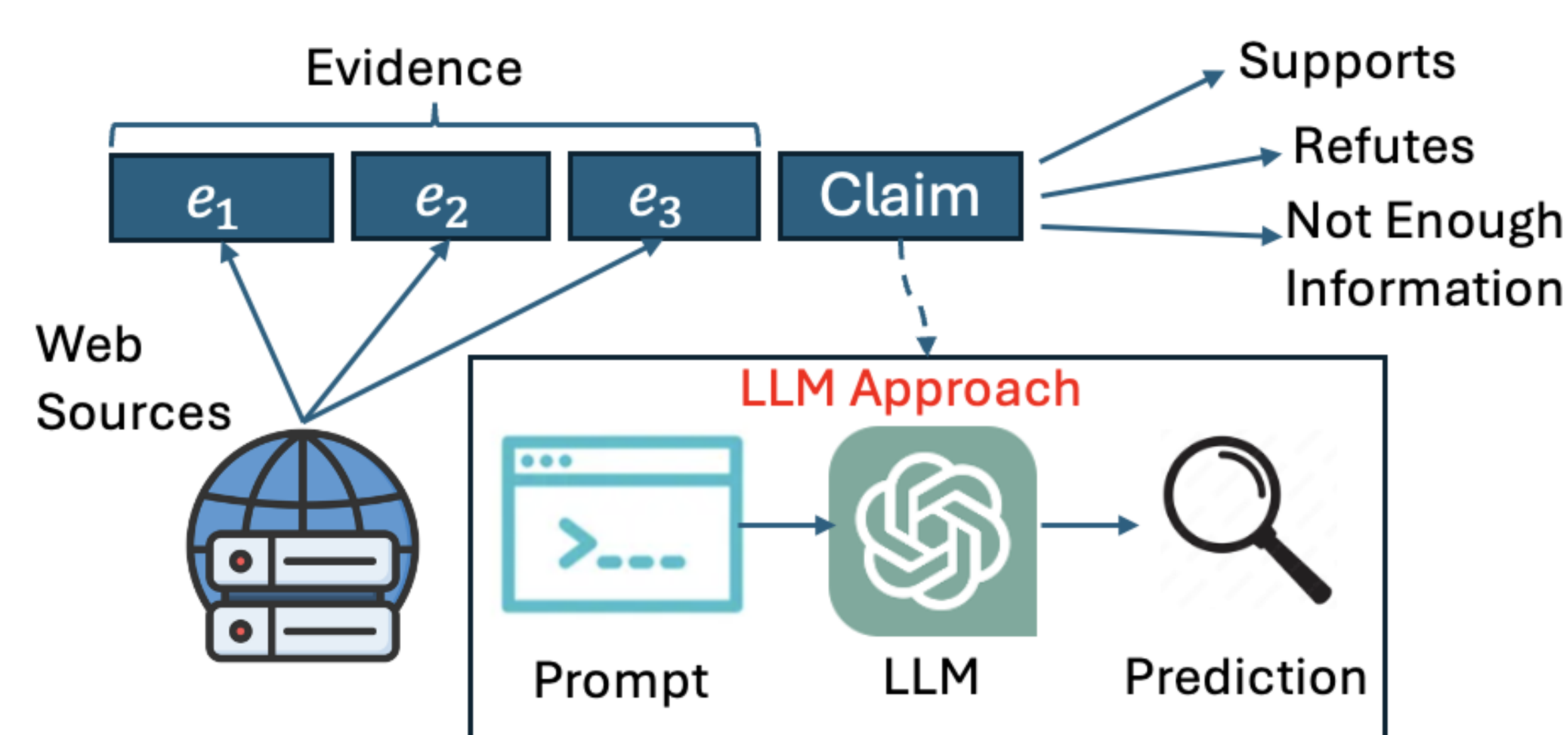


PolitiFact.comから抽出された誤った主張。

研究内容（方法・結果・結論）



自動ファクトチェック（AFC）の処理の流れ。



LLMを用いたプロンプト作成。

結果

Model	Evidence	$F1_{macro}$	$F1_{weighted}$	$F1_{micro}$	Acc.
Llama-3.2-3B-Instruct	No	0.624	0.624	0.624	0.504
	Yes	0.647	0.647	0.647	0.557
Llama-3.1-8B-Instruct	No	0.649	0.649	0.649	0.543
	Yes	0.668	0.668	0.668	0.589
Llama-3.1-70B-Instruct	No	0.689	0.689	0.689	0.684
	Yes	0.708	0.708	0.708	0.691
Llama-3.3-70B-Instruct	No	0.722	0.722	0.722	0.707
	Yes	0.747	0.747	0.747	0.734

Llama3.0ベースのLLMにおける二値分類の結果

結論

- 証拠検索手法を導入すると、LLMによるファクトチェックの精度を有意に向上させる。
- LLMをファクトチェックに利用できるが、判断根拠生成時にハルシネーションを起こす点に注意が必要。



連絡先：山岸研究室／国立情報学研究所 コンテンツ科学研究系

TEL：03-4212-2576 Email：yamagishilab-sec@nii.ac.jp