MAGIC: Multi-Argument Generation with Self-Refinement for Domain Generalization in Automatic Fact-Checking

Wei-Yu Kao and An-Zi Yen
Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
{wayner.cs09, azyen}@nycu.edu.tw

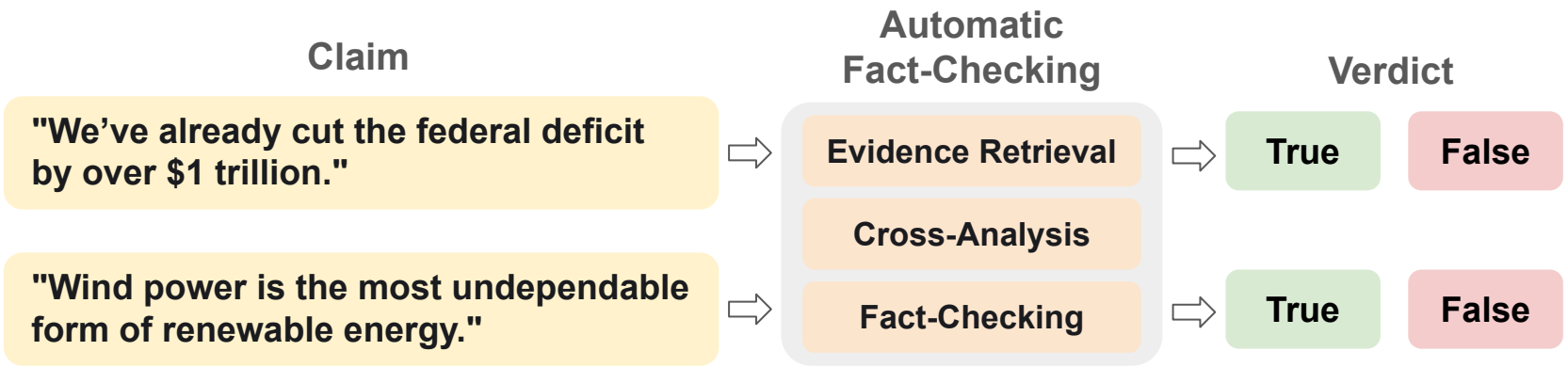
LREC-COLING 2024

NYCU NLP

Introduction

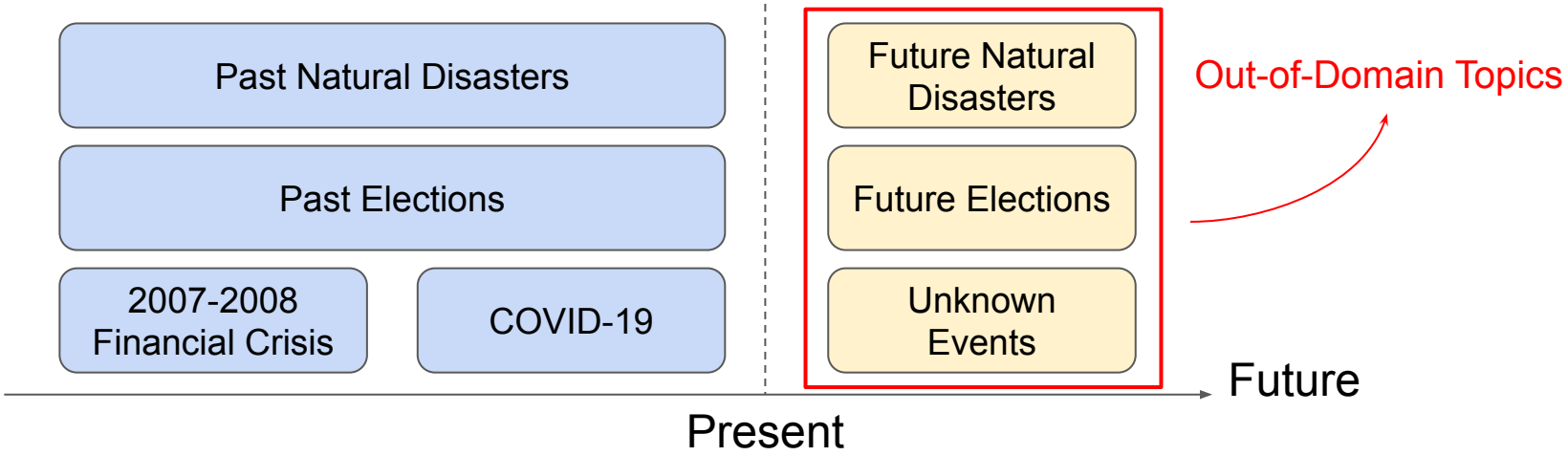
Automatic fact-checking is instrumental in validating the **abundance of content** on the internet, enabling individuals to access **unbiased** and **accurate information**.

In this paper, our focus is on enhancing **cross-domain** performance within automatic fact-checking by employing **multi-argument generation** techniques.



Significance of Domain Generalization in Fact-Checking

In their research, Khan et al. discovered the efficacy of automatic fact-checking **diminishes over time**, particularly when confronted with **emerging topics** such as **COVID-19** and **new election cycles**. These subjects introduce **out-of-domain data** for fact-checking models, leading to a decline in performance.



Khan et al. "WatClaimCheck: A new dataset for claim entailment and inference." ACL 2022.

Previous Study on Domain Generalization in Automatic Fact-Checking

Pan et al. tackled the issue by **enhancing fact-checking models** through the introduction of two methodologies: **Pretraining on Specialized Domains** and **Data Augmentation**.

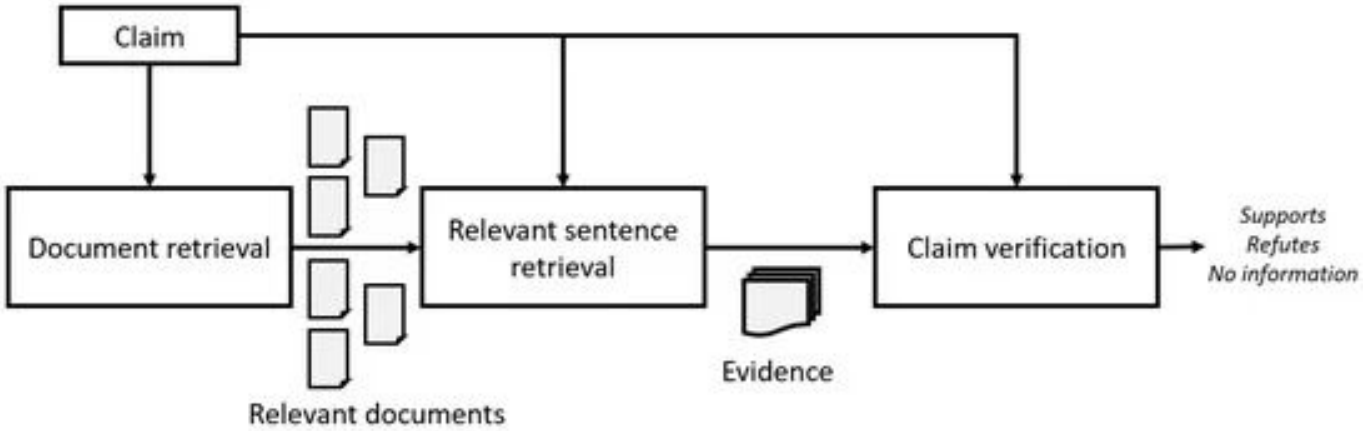
Pretraining on Specialized Domains may be inadequate for ambiguous domains **lacking pretrained models**, while Data Augmentation presents challenges concerning the **quality of generated claims**.

Pan et al. "Investigating Zero-and Few-shot Generalization in Fact Verification." AACL 2023.

Evidence-Based Automatic Fact-Checking

Research such as that conducted by Casillas, Ramón, et al., employs the **dense passage retrieval (DPR)** technique to retrieve **pertinent evidence**.

Incorporating this evidence into models enhances their ability to assess the authenticity of claims, thereby advancing the **effectiveness** of automatic fact-checking systems.



Casillas, Ramón, et al. "Automatic Fact Checking Using an Interpretable Bert-Based Architecture on COVID-19 Claims." Applied Sciences 2022.

The Influence of Evidence Granularity on Domain Generalization in Automatic Fact-Checking

Pan et al. conducted a study investigating the influence of varying levels of **evidence granularity** on the performance of fact-checking models with **out-of-domain** data.

Their findings suggest that employing **fine-grained evidence** can enhance the **domain generalization capability** of these models.

Evidence granularity example:

LREC-COLING 2024 will take place in Torino (Italy) on 20-25 May, 2024.

The conference will bring together practitioners in computational linguistics, speech, multimodality, and natural language processing.

Sentence-level evidences

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nullam faucibus mi quis velit. Morbi leo mi, nonummy eget tristique non, rhoncus non leo. Aliquam erat volutpat. Nullam lectus justo, vulputate eget mollis sed, tempor sed magna. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Ut enim ad minim veniam, quis nostrud exercitation ullamco.

Doc-level evidences

Pan et al. "Investigating Zero-and Few-shot Generalization in Fact Verification." AACL 2023.

MAGIC: Multi-Argument Generation and Self-Refinement

Multi-Argument Generation

- We employ **large language models** to generate arguments **assessing the authenticity** of claims based on **evidence** extracted from **each documents**.
- These arguments are **collectively evaluated** during the final fact-checking process.

$$A_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}; \mathcal{P}_{gen}) \quad (2)$$

$$\mathcal{A}_i = \{\langle A_{i,n} \rangle, \text{if}(\neg \text{drop}(A_{i,n}))\}_{n=1}^{N} \quad (3)$$

Self-Refinement

- We also implemented a **self-refinement mechanism** to ensure that the generated arguments are thoroughly **aligned** with the evidence.

MAGIC: Multi-Argument Generation and Self-Refinement

Self-Refinement

Building upon Madaan et al.'s research, we implemented a **self-refinement algorithm** tailored for **multi-argument generation**.

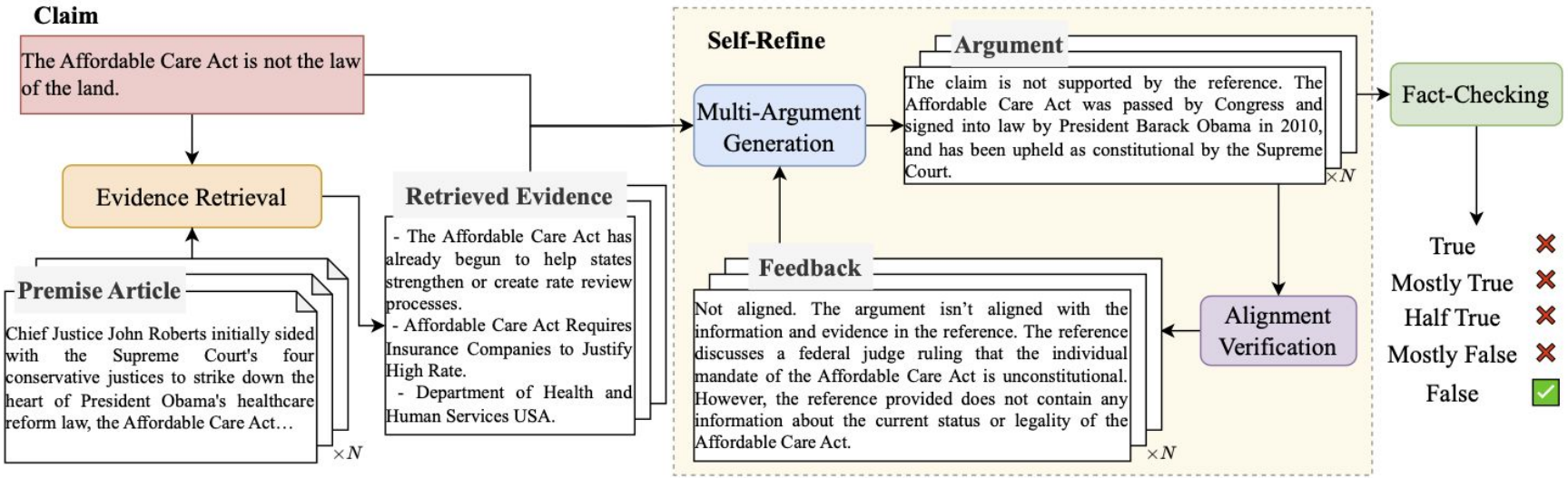
Our motivation stemmed from the observation that existing models occasionally **struggle** to generate arguments that **accurately align with the provided evidence**.

Algorithm 1 Self-refinement algorithm for multi-argument generation

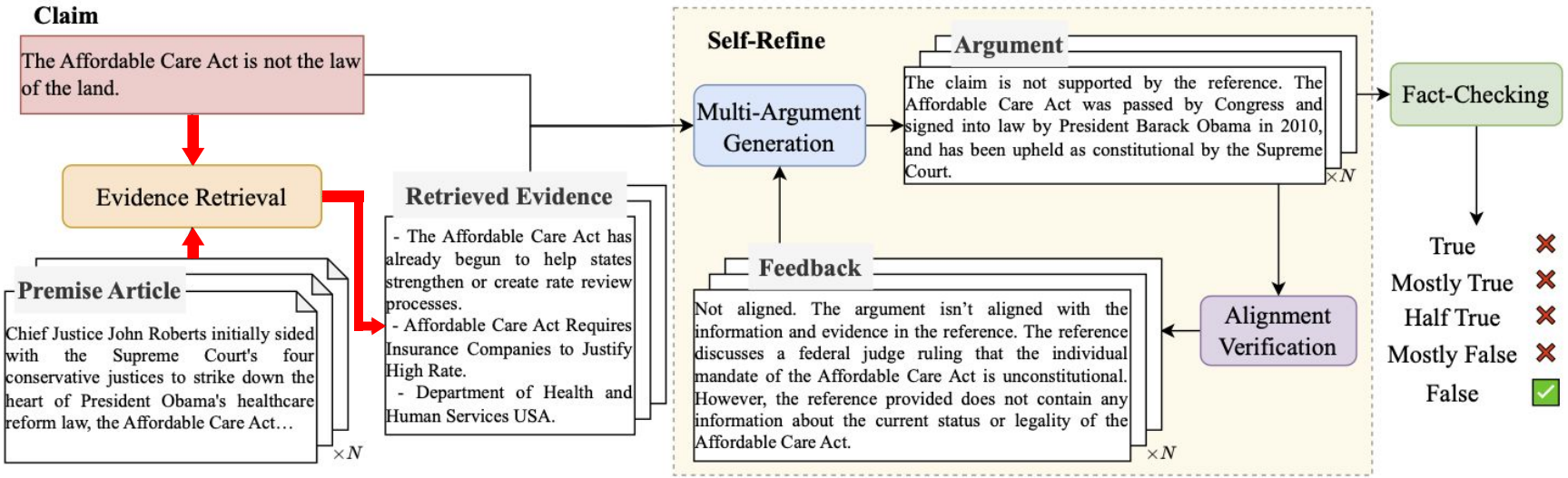
Require: Claim C_i, evidence $\mathcal{E}_{i,n}$, model $\mathcal{M}$, prompts $\{\mathcal{P}_{gen}, \mathcal{P}_f, \mathcal{P}_{rf}\}$, stop condition $\text{stop}_{align}(\cdot)$
$A^0_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}; \mathcal{P}_{gen})$
$f^0_{i,n} = \mathcal{M}(A^0_{i,n}, \mathcal{E}_{i,n}; \mathcal{P}_f)$
$t = 0$
$T = 10$
while not $\text{stop}_{align}(f^t_{i,n})$ **and** $t \leq T$ **do**
  $t = t + 1$
  $A^t_{i,n} = \mathcal{M}(C_i, \mathcal{E}_{i,n}, f^{t-1}_{i,n}; \mathcal{P}_{rf})$
  $f^t_{i,n} = \mathcal{M}(A^t_{i,n}, \mathcal{E}_{i,n}; \mathcal{P}_f)$
end while
return $A^t_{i,n}$

Madaan et al. "Self-refine: Iterative refinement with self-feedback." *Advances in Neural Information Processing Systems* 36 2024.

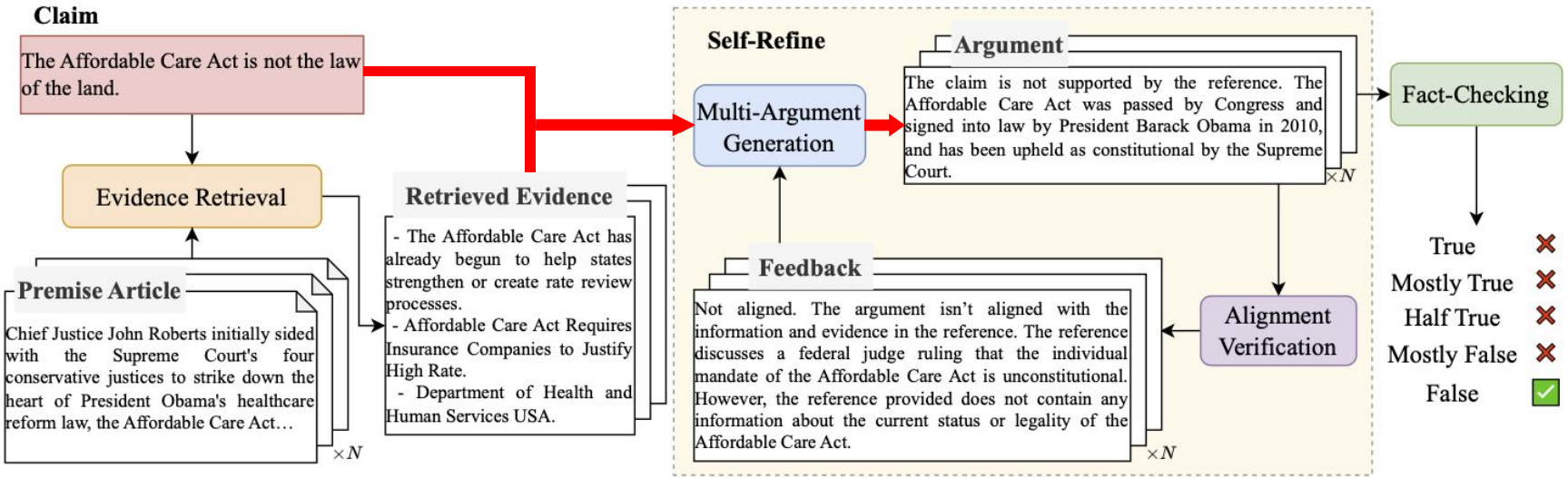
Our Framework



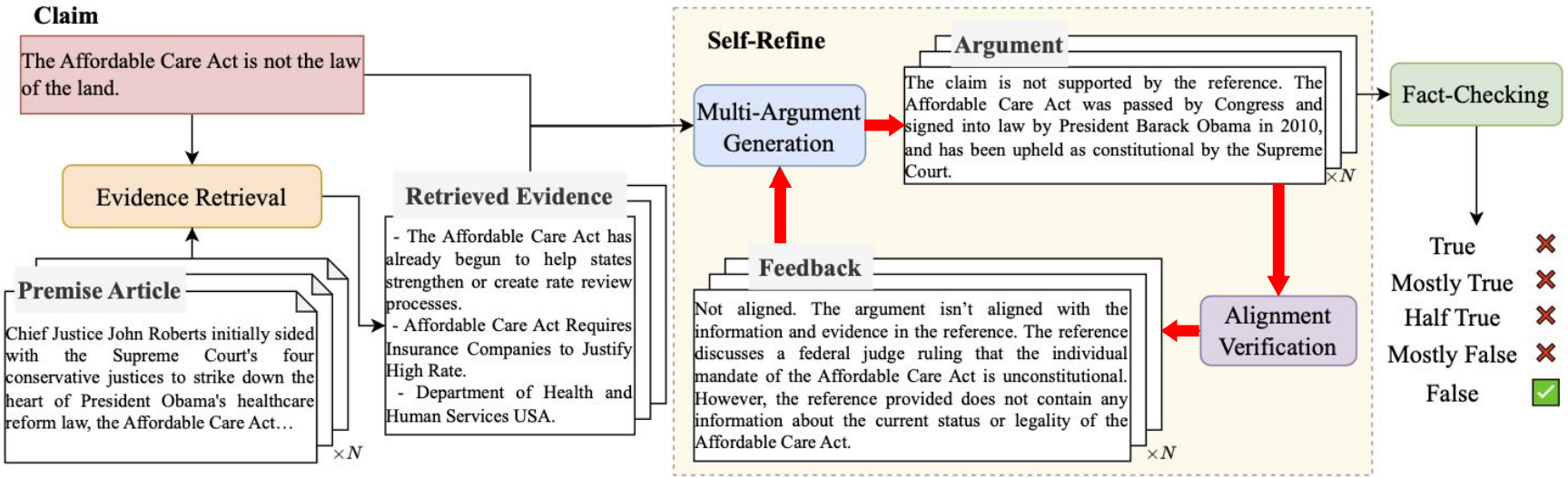
Our Framework



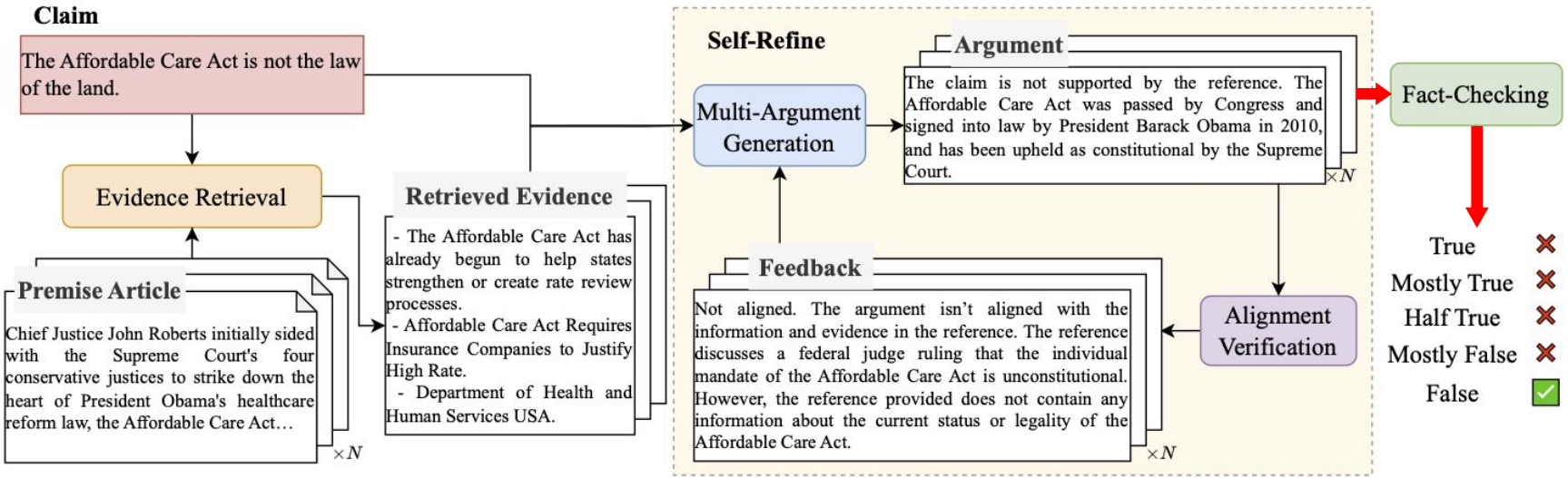
Our Framework



Our Framework



Our Framework



XClaimCheck Dataset

We gathered **16,177 claims** along with their **metadata** sourced from Khan et al.'s "WatClaimCheck", and augmented them as a **cross-domain dataset** encompassing **26 representative topics**, associated with 5 truth ratings.

To streamline our experiments, we organized these topics into **5 primary groups**.

Public policy and finance		Political issues		Legal and regulatory affairs		Infra. and services		Global affairs and security	
Topic	Count	Topic	Count	Topic	Count	Topic	Count	Topic	Count
Federal budget	824	Elections	1,167	Legal issues	554	Technology	145	Foreign policy	693
State budget	734	Candidate bio	801	LGBTQ	138	Energy	448	Immigration	983
Taxes	1,242	Jobs	914	Criminal justice	456	Transportation	267	Religion	235
Economy	1,337	Govt. regulation	248	Social sec.	168	Education	946	History	589
Health care	1,573			Homeland sec.	307	Sports	142	Military	420
Environment	436							Terrorism	410
Sum	6,146	Sum	3,130	Sum	1,623	Sum	1,948	Sum	3,330

Khan et al. "WatClaimCheck: A new dataset for claim entailment and inference." ACL 2022.

Experiment Setup

Baseline

- RoBERTa-Base
- Zero-shot LLMs: Vicuna-7b-v1.5

MAGIC

We used Vicuna-7b-v1.5 for mutli-argument generation and self-refinement.

- Encoder-Based Checker: RoBERTa-Base
- Seq2seq-Based Checker: Vicuna-7b-v1.5

RQ1: How effectively do small LLMs perform across different settings in our method?

Experiment Results

The evaluation metric is the **macro-averaged F-score**.

Overall Fact-Checking Performance:

- **MAGIC (encoder-based)** significantly **outperforms** the baseline RoBERTa and "MAGIC (seq2seq-based).
- **Self-refinement mechanism** within MAGIC shows that its inclusion clearly **benefits both** in-domain and out-of-domain fact-checking.

Model	Avg. / std.	In-domain	Out-of-domain
RoBERTa	0.2056 ± 0.0228	0.2307	0.1993
Zero-shot Vicuna	0.0667 ± 0.0000	0.0667	0.0667
MAGIC (seq2seq-based)	0.2049 ± 0.0156	0.2012	0.2058
w/o self-refine	0.1842 ± 0.0101	0.1846	0.1841
MAGIC (encoder-based)	**0.2500** ± 0.0175	**0.2661**	**0.2459**
w/o self-refine	0.2391 ± 0.0229	0.2459	0.2374

RQ2: How do larger LLMs like GPT-3.5 perform in our setup?

Experiment Results

Overall Fact-Checking Performance:

- **GPT-3.5-turbo** achieves competitive performance in identifying the veracity of **both in-domain and out-of-domain** claims.
- When paired with the **encoder-based checker**, it achieves the highest macro-averaged F-score in **in-domain** data.

Model	Avg. / std.	ID	OOD
MAGIC (encoder-based)	0.2500 ± 0.0175	0.2661	0.2459
Zero-shot GPT-3.5	0.2606 ± 0.0319	0.2505	0.2631
w/ encoder-based checker	0.2621 ± 0.0205	**0.2816**	0.2572
w/ seq2seq-based checker	**0.2623** ± 0.0250	0.2530	**0.2646**

RQ3: How do various models perform in identifying different claim ratings?

Experiment Results

MAGIC achieves promising results in assessing claims that fall into these **partially true or false** categories.

The ability to discern partially true or false claims is **crucial**, as most individuals can verify overtly false or true claims, but evaluating those that are **ambiguous** demands **greater expertise** and **information access**.

However, proposed method also shows **room for improvement** in discerning **"Mostly False"** and **"Half True"** claims.

	Rating											
	Pants on Fire		False		Mostly False		Half True		Mostly True		True	
Model	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD	ID	OOD
RoBERTa	**71.08%**	**64.15%**	**54.38%**	**52.26%**	17.19%	16.02%	17.51%	17.55%	5.80%	6.28%	27.68%	**35.65%**
MAGIC (encoder-based)	52.53%	45.13%	41.34%	35.44%	**28.38%**	**28.55%**	**26.25%**	**17.88%**	28.68%	22.77%	23.37%	28.34%
GPT-3.5 (seq2seq-based)	57.85%	52.31%	40.03%	39.02%	13.89%	14.06%	14.29%	13.10%	**35.33%**	**42.00%**	**27.83%**	30.05%

RQ4: How does MAGIC perform in the cross-domain fact-checking task?

Experiment Results

Best-performing model within each domain is **not necessarily** trained on that domain's data.

The degree of **relatedness** between the domains' subjects and the **complexity** of the issues impacts model performance:

- **"Public Policy and Finance"** has **significant correlations** with other subjects.
- **"Infrastructure and Services"** consists of subjects such as technology and energy, often utilizes **domain-specific terminology**, which makes them more **self-contained**.

	Test domain				
Training domain	Public Policy & Finance	Political Issues	Legal & Regulatory Affairs	Infra. & Services	Global Affairs & Security
Public Policy and Finance	**0.2809**	**0.2809**	0.2584	0.2403	**0.2641**
Political Issues	0.2272	0.2513	**0.2630**	0.2529	0.2414
Legal and Regulatory Affairs	0.2319	0.2471	0.2583	0.2289	0.2484
Infrastructure and Services	0.2503	0.2550	0.2360	**0.2964**	0.2338
Global Affairs and Security	0.2103	0.2348	0.2520	0.2616	0.2435

Conclusion and Future Work

Key results

- We introduce **cross-domain fact-checking** task, with **XClaimCheck dataset** and **MAGIC framework**.
- Experimental results demonstrate **effective enhancement** in fact-checking **out-of-domain** claims.

Future work

- An investigation spanning **diverse platforms and domains** is required to reflect the cross-domain fact-checking scenario.