Towards Automated Fact-Checking of Real-World Claims: Exploring Task Formulation and Assessment with LLMs

Premtim Sahitaj[1], Iffat Maab[2], Junichi Yamagishi[2],

Jawan Kolanowski, Sebastian Möller[1], and Vera Schmitt[1]

[1]Quality and Usability Lab
Technische Universität Berlin, Germany

[2]National Institute of Informatics (NII)
Tokyo, Japan

Motivation

- **Misinformation** spreads rapidly online, impacting public trust and decision-making.
- **Fact-Checking** is one strategy, with **pre-bunking** and **moderation** being complements.
- Manual fact-checking is slow and resource-intensive.
- **Evidence may change** over time and should be retrieved on demand.
- Need for automated, explainable verification systems → What about LLMs?

Introduction to **A**utomated **F**act-**C**hecking

AFC is the process of using computational techniques to assess the veracity of claims by retrieving relevant evidence and generating verdicts with supporting justifications

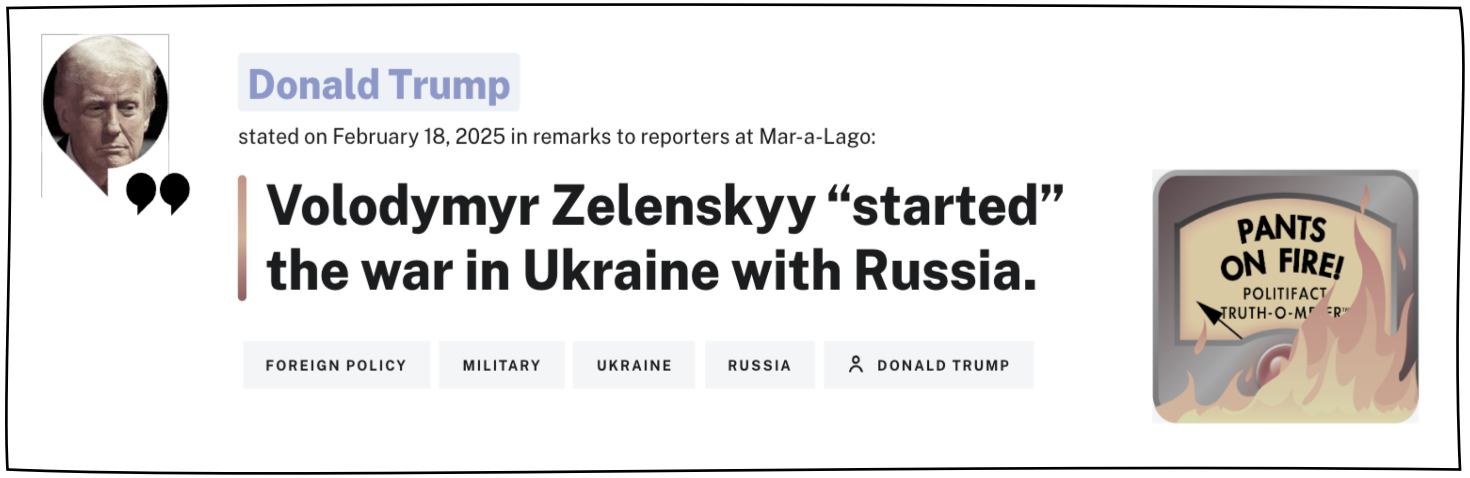
Guo et al. (2022) - A Survey on Automated Fact-Checking

Methodology

- **Data Collection**: Collect check-worthy claims and retrieve web evidence
- **Task Formulation:** Consider AFC as a multi-component task with three objectives: (1) step-by-step analysis, (2) verdict prediction, and (3) justification generation.
- **Label Schemes:** Evaluate different granularity levels to understand the impact on task performance.
- **Model Evaluation:** Compare performance across various LLM sizes (3B, 8B, 70B) in a few-shot inference setting.
  - Assess classification accuracy and justification quality using a reference-free metric.
  - Evaluate with and without evidence.

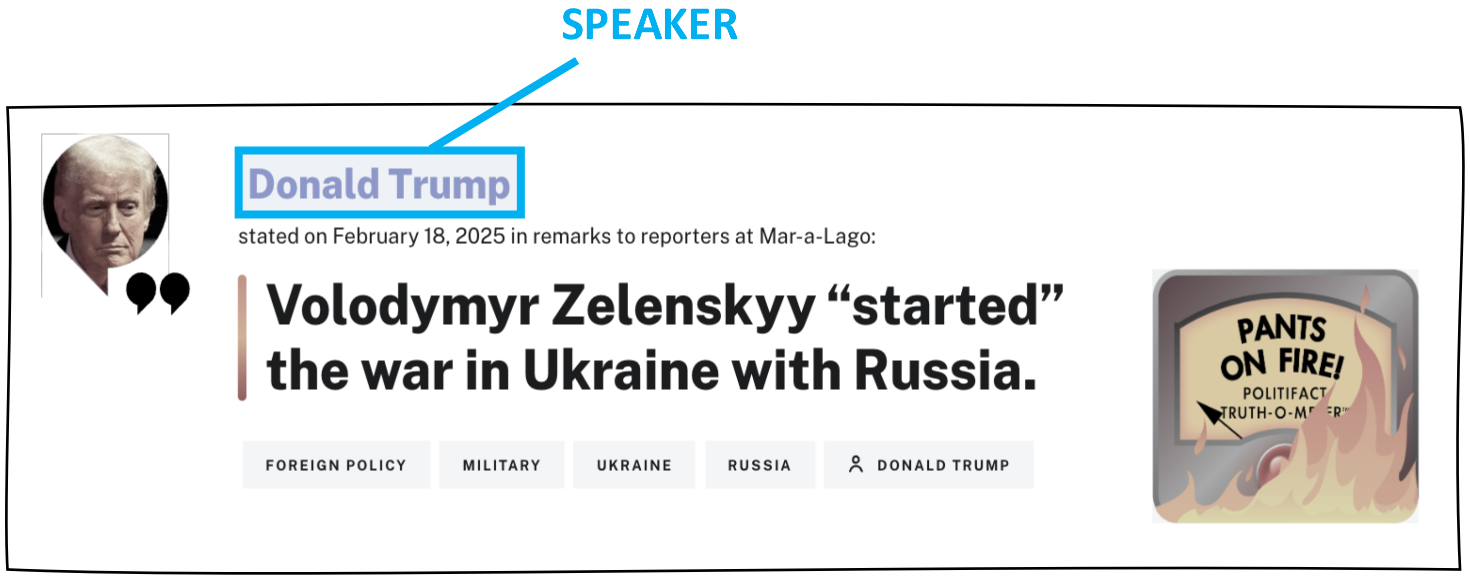
Data Collection

- Assumption: Fact-Checking experts can accurately identify check-worthy claims
- Data collected from **PolitiFact** (2007–2024) containing 17,856 claims made by public speakers



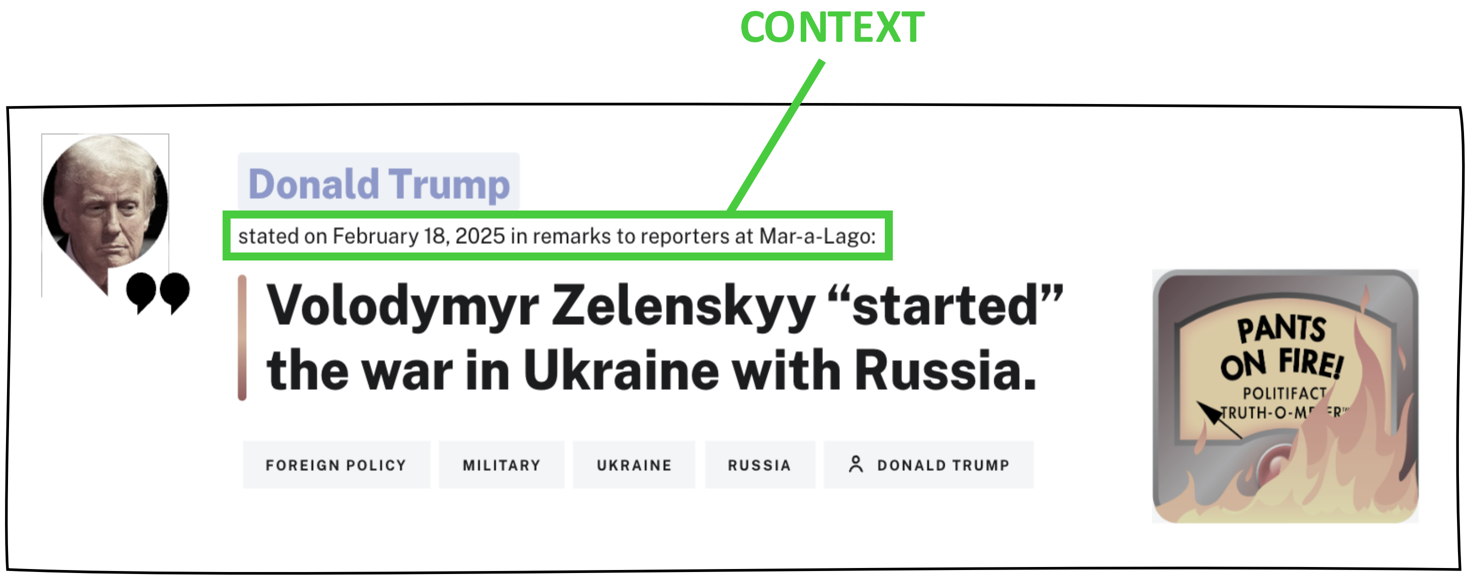
Data Collection

- Assumption: Fact-Checking experts can accurately identify check-worthy claims
- Data collected from **PolitiFact** (2007–2024) containing 17,856 claims made by public speakers



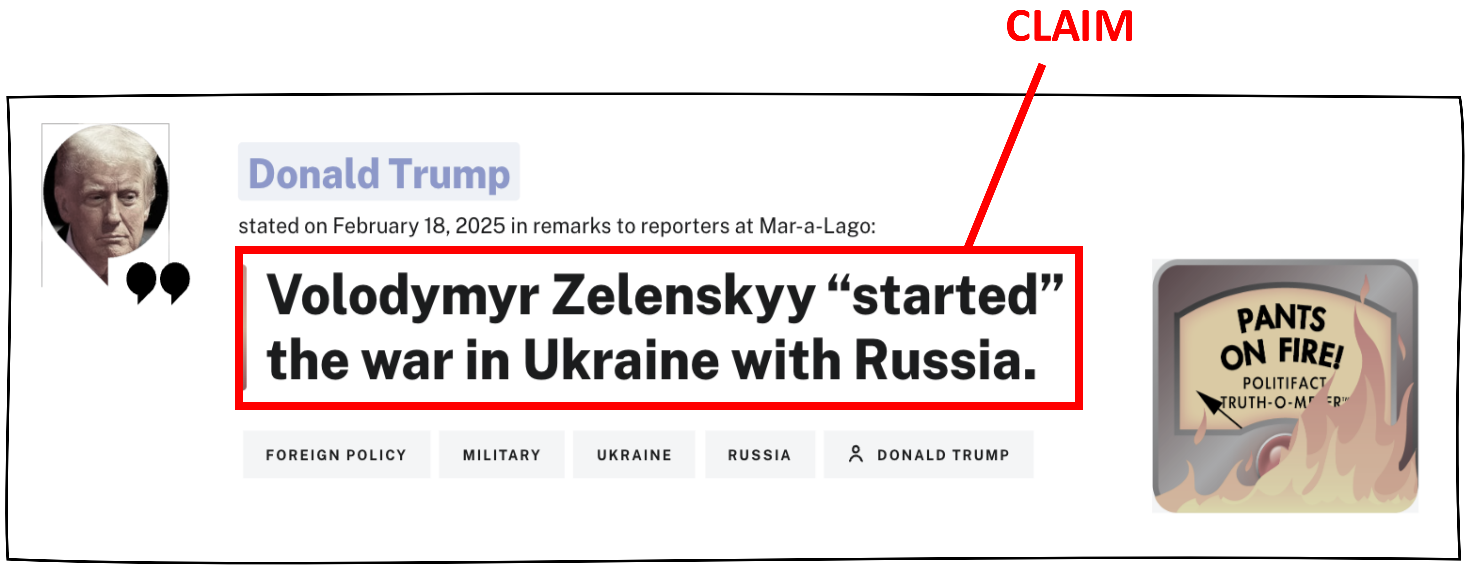
Data Collection

- Assumption: Fact-Checking experts can accurately identify check-worthy claims
- Data collected from **PolitiFact** (2007–2024) containing 17,856 claims made by public speakers



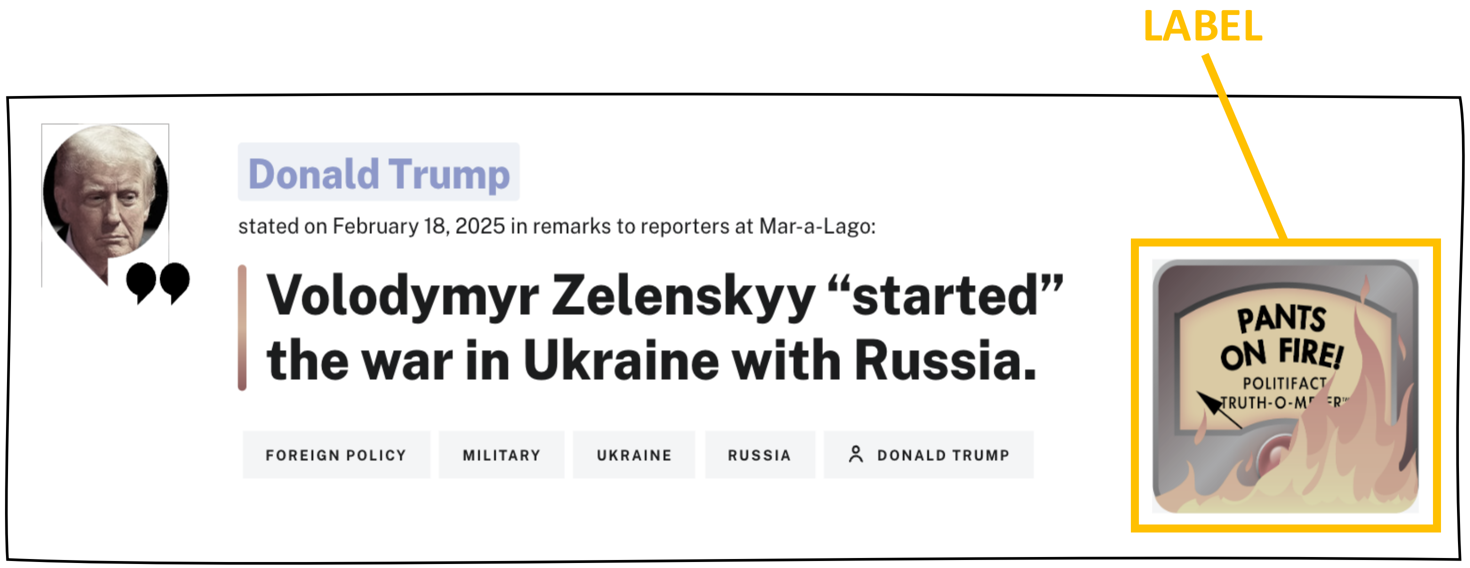
Data Collection

- Assumption: Fact-Checking experts can accurately identify check-worthy claims
- Data collected from **PolitiFact** (2007–2024) containing 17,856 claims made by public speakers



Data Collection

- Assumption: Fact-Checking experts can accurately identify check-worthy claims
- Data collected from **PolitiFact** (2007–2024) containing 17,856 claims made by public speakers



Data Collection - Labels

Table 1
Definitions of the original PolitiFact rating system labels.

Label	Definition
TRUE	... is accurate and there's nothing significant missing.
MOSTLY TRUE	... is accurate but needs clarification or additional information.
HALF TRUE	... is partially accurate but leaves out important details or takes things out of context.
MOSTLY FALSE	... contains an element of truth but ignores critical facts [...].
FALSE	... is not accurate.
PANTS ON FIRE	... is not accurate (thus false) **and** makes a ridiculous claim.

Data Collection - Labels

Table 1
Definitions of the original PolitiFact rating system labels.

Label	Definition
TRUE	... is accurate and there's nothing significant missing.
MOSTLY TRUE	... is accurate but needs clarification or additional information.
HALF TRUE	... is partially accurate but leaves out important details or takes things out of context.
MOSTLY FALSE	... contains an element of truth but ignores critical facts [...].
FALSE	... is not accurate.
~~PANTS ON FIRE~~	... is not accurate (thus false) **and** makes a ridiculous claim.

Label Schemes

Table 2
Distribution of data for five, three, and two-class Settings.

Per-class Labels	5-class Count	5-class Percentage
true	2531	14.18%
mostly-true	3347	18.75%
half-true	3534	19.79%
mostly-false	3212	17.99%
false	5231	29.30%

Label Schemes

Table 2
Distribution of data for five, three, and two-class Settings.

Per-class	**5-class**		**3-class**	
Labels	**Count**	**Percentage**	**Count**	**Percentage**
true	2531	14.18%	-	-
mostly-true	3347	18.75%	5878	32.92%
half-true	3534	19.79%	3534	19.79%
mostly-false	3212	17.99%	8443	47.29%
false	5231	29.30%	-	-

Label Schemes

Table 2
Distribution of data for five, three, and two-class Settings.

Per-class	**5-class**		**3-class**		**2-class**	
Labels	**Count**	**Percentage**	**Count**	**Percentage**	**Count**	**Percentage**
true	2531	14.18%	-	-	-	-
mostly-true	3347	18.75%	5878	32.92%	9412	52.71%
half-true	3534	19.79%	3534	19.79%	-	-
mostly-false	3212	17.99%	8443	47.29%	8443	47.29%
false	5231	29.30%	-	-	-	-

Evidence Retrieval

- **Evidence retrieval** grounds fact verification by providing external context.
- Generally, we have the issue that **credible information** is not readily available.
- We use a **restricted** web search to retrieve the **top 10** results for each data point as evidence:
  - **Exclude** common **fact-checking sites** to avoid explicit leaks from previous verification efforts
  - **Exclude** a set of **terms** related to fact-checking
  - Retrieve title, snippet, URL, and date.
- We do not utilize search query optimization!

Task Formulation

- Define fact-checking as a multi-component task with three objectives:
  - **Step-by-Step Analysis**: Perform detailed analysis over the claim and evidence if available.
  - **Verdict Prediction**: Assign a veracity label based on analysis.
  - **Justification Generation**: Provide a concise explanation that justifies the verdict.

SYSTEM: You are an intelligent decision support system for automated fact-checking.
Your tasks are:

1. Analyze the claim step-by-step.
2. Classify the claim's veracity based on your analysis. [LABELS]
3. Provide a concise natural language explanation for the verdict prediction.

USER: [SPEAKER][CONTEXT] the claim [CLAIM]. Evidence: [EVIDENCE]

Example

Source: New York Times Editorial Board
Background: The editorial board is made up of 16 journalists ...
Context: stated on June 14, 2017 in a New York Times editorial
Claim: ”A political map circulated by Sarah Palin’s 2019s PAC incited Rep. Gabby Giffords’s 2019 shooting”
Label: False

Example

Source: New York Times Editorial Board
Background: The editorial board is made up of 16 journalists ...
Context: stated on June 14, 2017 in a New York Times editorial
Claim: "A political map circulated by Sarah Palin's 2019s PAC incited Rep. Gabby Giffords's 2019 shooting"
Label: False

Table 3
Snippets of web evidence related to Sarah Palin and the New York Times editorial case.

Date	Snippet
Feb 10, 2022	Sarah Palin testified Thursday that she felt "mortified" ...
Sep 10, 2020	Palin's political action committee circulated a map of ...
Aug 16, 2017	The Times subsequently issued a correction stating that no such link ...

Example

Source: New York Times Editorial Board
Background: The editorial board is made up of 16 journalists ...
Context: stated on June 14, 2017 in a New York Times editorial
Claim: "A political map circulated by Sarah Palin's 2019s PAC incited Rep. Gabby Giffords's 2019 shooting"
Label: False

Table 3
Snippets of web evidence related to Sarah Palin and the New York Times editorial case.

Date	Snippet
Feb 10, 2022	Sarah Palin testified Thursday that she felt "mortified" ...
Sep 10, 2020	Palin's political action committee circulated a map of ...
Aug 16, 2017	The Times subsequently issued a correction stating that no such link ...

Section	Content
Reasoning	The New York Times editorial board claimed that a political map circulated by Sarah Palin's PAC incited Rep. Gabby Giffords' shooting. However, there is no established link between the two events, and the Times subsequently issued a correction. The case was later brought to court, with Sarah Palin suing the New York Times for defamation. The court ultimately ruled in favor of the New York Times, stating that the editorial did not defame Palin.
Verdict	False
Explanation	The statement is not accurate. There is no established link between the political map circulated by Sarah Palin's PAC and the shooting of Rep. Gabby Giffords, and the New York Times issued a correction after publishing the claim.

Figure 2: Analysis of the New York Times editorial case involving Sarah Palin.

Experimental Setup

- Few-shot inference with synthetic examples, one per label.
- Models Evaluated:
  - Llama-3.2-3B,
  - Llama-3.1-8B,
  - Llama-3.1-70B,
  - Llama-3.3-70B (distilled from Llama-3.1-405B)
- Evaluation Metrics:
  - F1 score for verdict prediction
  - TIGERScore for justification quality
- Comparison Variables:
  - With vs. without evidence integration.
  - Impact of label complexity on performance.

Hypotheses

H_1: Classification performance **decreases** as label complexity **increases**.

H_2: Justification quality **decreases** as label complexity **increases**.

H_3: Incorporating external evidence **improves** both classification accuracy and justification quality.

H_4: Larger models **perform better** in classification tasks and produce higher-quality justifications.

H_5: Smaller models **benefit more** from evidence integration.

Results

Table 7
Aggregated Results.

Model	**Evidence**	**5-Class** $F1_{micro}$	**5-Class** TIGER	**3-Class** $F1_{micro}$	**3-Class** TIGER	**Binary** $F1_{micro}$	**Binary** TIGER
Baseline	-	0.213	-	0.371	-	0.501	-
3.2-3B-Instruct	No	0.273	-3.995	0.464	-4.069	0.624	-3.870
	Yes	0.321	-3.116	0.498	-3.205	0.647	-3.150
3.1-8B-Instruct	No	0.293	-3.416	0.472	-3.391	0.649	-3.367
	Yes	0.339	-2.578	0.525	-2.751	0.668	-2.741
3.1-70B-Instruct	No	0.356	-2.554	0.542	-2.610	0.689	-2.560
	Yes	0.389	-2.466	0.556	-2.524	0.708	-2.433
3.3-70B-Instruct	No	0.357	-2.361	0.556	-2.383	0.722	-2.303
	Yes	**0.405**	**-1.686**	**0.589**	**-1.884**	**0.747**	**-1.739**

Hypotheses

H_1: Classification performance decreases as label complexity increases.

H_2: Justification quality decreases as label complexity increases.*

H_3: Incorporating external evidence improves both classification accuracy and justification quality.

H_4: Larger models perform better in classification tasks and produce higher-quality justifications.

H_5: Smaller models benefit more from evidence integration.*

*rejected

Conclusions & Future Works

- LLMs demonstrate utility for automated fact-checking
- Integrating web evidence does improve task performance
- Evaluation of LLMs is difficult (e.g., parametric vs. contextual knowledge)
- Truthfulness label can be ambiguous between annotators, investigate alternative schemes
  → e.g., FEVER-style

- **Future Work**:
  - **Knowledge Base Construction** from Fact-Checking Articles for Claim Matching and Adaption
  - **Apply** more sophisticated RAG approaches for web evidence, e.g. QA, Chain of RAG,
  - **Deployment** and evaluation as component in real-world user scenarios
  - **Comparison** against community-driven approaches.

Thank your for your attention!

Do you have any questions?

Mail: sahitaj@tu-berlin.de