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

Premtim Sahitaj[1,2,*], Iffat Maab[3], Junichi Yamagishi[3], Jawan Kolanowski[4], Sebastian Möller[1,2] and Vera Schmitt[1,2]

[1]Technische Universität Berlin, Quality and Usability Lab, Berlin, 10587, Germany
[2]Deutsches Forschungszentrum für Künstliche Intelligenz, Speech and Language Technology Lab, Berlin, 10559, Germany
[3]National Institute of Informatics, Digital Content and Media Sciences Research Division, Tokyo, 101-8430, Japan
[4]Harz University of Applied Sciences, Faculty of Automation and Computer Science, Wernigerode, 38855, Germany

Abstract

Fact-checking is necessary to address the increasing volume of misinformation. Traditional fact-checking relies on manual analysis to verify claims, but it is slow and resource-intensive. This study establishes baseline comparisons for Automated Fact-Checking (AFC) using Large Language Models (LLMs) across multiple labeling schemes (binary, three-class, five-class) and extends traditional claim verification by incorporating analysis, verdict classification, and explanation in a structured setup to provide comprehensive justifications for real-world claims. We evaluate Llama-3 models of varying sizes (3B, 8B, 70B) on 17,856 claims collected from PolitiFact (2007–2024) using evidence retrieved via restricted web searches. We utilize TIGERScore as a reference-free evaluation metric to score the justifications. Our results show that larger LLMs consistently outperform smaller LLMs in classification accuracy and justification quality without fine-tuning. We find that smaller LLMs in a few-shot inference scenario provide comparable task performance to fine-tuned Small Language Models (SLMs) with large context sizes, while larger LLMs consistently surpass them. Evidence integration improves performance across all models, with larger LLMs benefiting most. Distinguishing between nuanced labels remains challenging, emphasizing the need for further exploration of labeling schemes and alignment with evidences. Our findings demonstrate the potential of retrieval-augmented AFC with LLMs.

Keywords

Automated Fact-Checking, Large Language Models, Retrieval-Augmented Generation,

1. Introduction

Misinformation, whether spread inadvertently or with the intention to deceive, is a global challenge that can be mitigated effectively through fact-checking efforts [1]. Generally, fact-checking is defined as the assessment of the truthfulness of a check-worthy claim [2, 3]. For fact-checking to be effective, fact-checking itself must be convincing and justified [4]. A well-known source of human-verified knowledge is PolitiFact[1], where experts manually identify check-worthy claims from news and social media and document their verification efforts in written articles. Traditional fact-checking of these claims relies on human-driven exploration, analysis, and conclusion. Consequently, this process is rather slow and expensive, lagging behind the rapid spread of misinformation. Delayed fact-checking efforts allow false narratives to take hold, distort reality, and influence public opinion, a vulnerability that is often exploited by bad actors [5]. Additionally, moderation policies [6] and pre-bunking methodologies [7] offer proactive strategies by addressing misinformation before it spreads widely.
AFC systems assist human efforts to combat misinformation by leveraging state-of-the-art techniques from areas such as Natural Language Processing (NLP), Natural Language Generation (NLG), and Information Retrieval (IR). Ideally, these systems automatically extract claims from the presented media,

ROMCIR 2025: The 5th Workshop on Reducing Online Misinformation through Credible Information Retrieval (held as part of ECIR 2025: the 47th European Conference on Information Retrieval), April 10, 2025, Lucca, Italy
*Corresponding author.
sahitaj@tu-berlin.de (P. Sahitaj); maab@nii.ac.jp (I. Maab); jyamagis@nii.ac.jp (J. Yamagishi); u37871@hs-harz.de (J. Kolanowski); sebastian.moeller@tu-berlin.de (S. Möller); vera.schmitt@tu-berlin.de (V. Schmitt)
0000-0003-3908-5681 (P. Sahitaj); 0000-0003-3057-0760 (S. Möller); 0000-0002-9735-6956 (V. Schmitt)

[1]https://politifact.com

retrieve relevant and credible references, and provide evidence-based verdicts on the aggregated results. As opposed to style-based detection approaches that learn to distinguish claims based on writing patterns, AFC systems follow a knowledge-based approach that relies on verification knowledge to make judgements on claims [8]. Expert fact checkers can utilize AFC systems as intelligent decision support assistance to eliminate repetitive manual tasks, highlight inconsistencies, and present their findings [9].

Humans often distrust fact-checking work that challenges their beliefs, perceiving them as biased or manipulated [10]. This skepticism is likely to be aggravated with closed systems, where the lack of transparency around internal mechanisms and design decisions further erodes trust. Brandtzaeg and Følstad [10] argue that to strengthen trust, fact-checking processes must be made transparent.

LLMs such as GPT-4, Claude 3.5 Sonnet, and Llama-3, have provided significant potential for a broad range of text-to-text reasoning tasks. Integrating LLMs as an inference engine into AFC systems may enhance transparency by generating veracity predictions and the accompanying natural language explanations. However, Setty [11] demonstrates that, for AFC-related classification tasks, fine-tuned small language models (SLM) outperform LLMs. This indicates that further research is needed to effectively utilize LLMs for AFC.

This paper investigates the task formulation and assessment for AFC of real-world claims with LLMs to establish baselines in various settings and to evaluate whether truthfulness ratings can be effectively modeled or if alternative approaches to claim annotations and task formulation are necessary. Based on these findings, future approaches can make more informed design choices and improve the reliability and effectiveness of AFC systems. We propose a framework for AFC with LLMs in a few-shot setup without model fine-tuning for claim analysis, claim veracity prediction, and the generation of justifications as natural language explanations. In the scope of this work, we assess the performance of our framework on 17,856 real-world claims from PolitiFact based on three labeling schemes, with or without web evidence, and across models of different sizes (i.e. 3B, 8B, 70B). Using reference-free evaluation metrics and conducting extensive experiments, we provide insight into how evidence integration, model size, and labeling complexity impact system performance. Additionally, we consider fine-tuning small state-of-the-art classification models for estimating the upper bound of predictive performance extractable from different components of the data points in the collected dataset and assess the performance of LLMs with a few-shots relative to this limit. Thus, our findings contribute to the development of more robust and transparent AFC systems using LLMs.

2. Related Work

Our work builds on the existing body of research in fact-checking and retrieval-augmented generation while addressing several gaps in the literature. Prior studies have established the value of transformer-based architectures such as BERT and GPT for tasks ranging from sequence classification to text generation [12, 13] and have shown that integrating retrieval mechanisms via RAG can improve factual grounding [14]. Automated fact-checking frameworks typically consist of claim detection, evidence retrieval, and claim verification [4]. Claim detection identifies check-worthy claims [3], often guided by factors such as relevance or harm, while evidence retrieval involves collecting and selecting relevant information to justify verdicts [15]. Claim verification can be broken down into two main tasks: (a) verdict prediction and (b) justification production [4].

Our approach unifies the components of claim verification into a single structured framework. However, unlike previous works that often rely on fine-tuned models or separate stages for classification and explanation [16, 17, 18, 19, 20], we propose a few-shot inference setup using LLMs that simultaneously produces analysis, verdict classification, and justification generation in a structured format. Prior work has highlighted that while LLMs can generate justifications, they are prone to hallucinations [21] and may lead users to over-rely on potentially incorrect explanations [22]. Our integrated approach is motivated by chain-of-thought reasoning techniques [23], which implement step-by-step analysis and aims to facilitate consistency between the generated verdict and its justification.

Additionally, we evaluate our approach on open-source LLMs of different scale, in contrast to similar prior work that utilizes only closed-source models such as ChatGPT [19]. Moreover, our experimental analysis extends previous findings by exploring the effects of varying label granularity, from binary to multi-class setups, and by systematically investigating the impact of evidence integration on both classification performance and justification quality. While some works, for example, Augenstein et al. [24], have studied diverse labeling schemes, our study directly compares performance across a hierarchy of related schemes. This empirical insight addresses a notable gap in the literature regarding the interplay between label complexity, model scale, and the integration of external evidence.

3. Dataset

Fact-checking organizations, that document their efforts and share them publicly, offer a great opportunity to analyze relevant misinformation and model the verification process. Moreover, by providing the initial judgment on what is check-worthy or not, fact-checking experts greatly reduce the complexity of the task at hand [2]. At PolitiFact, experts select check-worthy claims by determining whether they are verifiable as opposed to opinions and personal experiences, potentially misleading, significant enough to influence public discourse, likely to be repeated, or if a typical reader would reasonably question their truthfulness. The content at PolitiFact is localized around topics that can be found in US news. In this work, we utilize a dataset collected from PolitiFact's online repository of fact-checking efforts. PolitiFact is a frequently used source of misinformation data, as seen in LIAR LIAR [25] or Mocheg [26]. We collect 23,495 data points from English PolitiFact articles between 2007 and the January 26, 2024. Claims not attributed to public figures (i.e. social media posts) were excluded, as these were predominantly evaluated as fake, resulting in a refined dataset of 17,856 claims. In the context of this research, we are interested in collecting the claims that have been deemed check-worthy, the entity that shared said claim, the context in which the claim has been produced, and finally the rating that has been assigned to the claim. We also match and provide the background descriptions of the entity that produced the claim. Figure 1 illustrates the available features.

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

Figure 1: Example data point of a statement made by the New York Times Editorial Board and evaluated by PolitiFact as False.

PolitiFact's rating system follows an ordinal six-class labeling scheme. Table 1 provides the official descriptions of these six classes.

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.

While PolitiFact assigns a separate *PANTS ON FIRE* label to document the characteristic of ridiculousness in claims, we are only interested in the dimension of truthfulness and therefore treat this special label as a sub-case of *False*. Thus, we merge the classes, discard the sixth label, and reduce the overall set of labels to five. Table 2 illustrates the resulting distribution of classes.

4. Methodology

This Section outlines the methodology used to design and evaluate our framework for automated fact-checking with LLMs. Following the description of the data collection from PolitiFact, we formulate the problem and the experimental setup. Specifically, we discuss model selection, labeling scheme choices, and evidence retrieval.

4.1. Task Formulation

The approach in this study is motivated by the need to enhance coherence, consistency, and interpretability in automated fact-checking systems. By combining reasoning, classification, and explanation as justification within a single framework, we aim to leverage intermediate analysis to improve performance and ensure consistency between outputs. This study approaches automated fact-checking as a multi-component task with three key objectives:

1. **Reasoning**: Producing a detailed, step-by-step analysis of the claim using the available information.
2. **Verdict**: Assigning a veracity label to the claim based on a predefined set of categories.
3. **Explanation**: Providing a clear and concise explanation in natural language to support the assigned verdict.

The reasoning task follows the idea of chain-of-thought reasoning [23] by constructing a step-by-step analysis of the available information as a natural language explanation [27]. Thus, the verdict classification is integrated with both preceding analysis and subsequent explanation to enhance performance, building on insights from existing research. Zhang et al. [28] demonstrate that jointly generating explanations and predictions outperforms explain-then-predict models. Similarly, Atanasova et al. [29] find that generating fact-checking explanations alongside veracity predictions improves both the performance and the quality of the explanations. These tasks are addressed within a few-shot classification framework, utilizing instruction-based prompts to guide LLMs in generating structured outputs.

4.2. Prompt Design

We design the prompts based on the previously outlined problem formulation and established principles of prompt engineering [30]. Each prompt is composed of three main components: system, user, and assistant. The system message sets the model's context and provides the instructions, including the selected labeling scheme. The user message specifies the speaker, context, and claim, with evidence included when available. The assistant message contains the model's response to the input. To simulate a chat history with desired outputs, few-shot examples, one per label, are included as user and assistant message pairs following the system message and preceding the actual input.

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]

ASSISTANT:

To ensure consistency and enable automated processing, we enforce a structured output format using the `vLLM`[2] and `outlines`[3] libraries. In this context, structure refers to the property of generated output satisfying a constrained syntax [31]. The output is generated as a parsable JSON object with the following properties: reasoning, verdict, and explanation. Reasoning as free-text, step-by-step analysis of the claim. The verdict as the predicted veracity label, constrained to any option of the predefined set of labels. Lastly, the concise natural language explanation arguing the verdict prediction.

4.3. Model Selection

To evaluate performance across different model scales, we selected a range of LLMs from the Llama 3 series. We choose Llama architecture models due to their state-of-the-art performance and open-source availability, making them well-suited for evaluating automated fact-checking systems. The models used in this study are Llama-3.2-3B, Llama-3.1-8B, Llama-3.1-70B, Llama-3.3-70B in their instruction-finetuned state. The selection covers varying parameter sizes (3B, 8B, 70B) to investigate the relationship between model scale and task performance. Our strategy is to evaluate the most recent model available at each size. The 3.2 line was the first to introduce the 3B size, while the only 8B version is found in the 3.1 line. For the 70B size, checkpoints are available in both the 3.1 and 3.3 lines. All models have a December 2023 knowledge cutoff. During pre-training, the 3.2 models processed 9 trillion tokens, whereas the 3.1 and 3.3 models processed 15 trillion tokens. The 3.3 70B Llama model achieves comparable performance to the 3.1 405B model[4], making it one of the most performant open source models at this size. This justifies its inclusion as an additional option in model selection. All models are used in their instruction-tuned state to ensure alignment with the task. Instead of further fine-tuning, we rely on the models' available capabilities to perform few-shot reasoning, classification and explanation.

4.4. Label Schemes

Fact-checkers adopt varied approaches to labeling schemes, reflecting different priorities and methodologies. Some, such as FullFact[5], rely solely on justifications without assigning explicit ratings to claims. Others, like PolitiFact and Snopes[6], implement labeling systems grounded in the idea of truthfulness. A further extension of these schemes includes labels for scenarios where evidence is incomplete or unavailable. In the AFC community, truthfulness labels are frequently mapped to a conceptual dimension that evaluates factuality based on available ground-truth evidence. Labels such as supported, refuted, cherry-picked, or not enough information (NEI) are commonly used [32, 33, 26], requiring significant human effort for exploration and annotation. While these approaches provide valuable insights, they also introduce complexities related to interpretation and consistency in annotations. We postpone this perspective to future work. Our focus in this study is to assess whether fact-checking can be effectively modeled across different granularities of truthfulness on the collected data. Specifically, we aim to evaluate the trade-offs between simpler and more nuanced labeling schemes in terms of their impact on classification performance and justification quality. To evaluate the impact of label granularity on fact-checking performance, we merge the five original PolitiFact labels (*True*, *Mostly True*, *Half True*, *Mostly False*, *False*) into coarser schemes, progressively reducing complexity while preserving interpretability. In the three-class scheme, the original labels true and false are grouped into mostly true and mostly false, respectively. In the binary scheme, the label half-true is merged into mostly true. We aim to align our label aggregation with PolitiFact's definitions, as introduced in Table 1. Table 2 illustrates the resulting distributions.

[2]https://github.com/vllm-project/vllm
[3]https://github.com/dottxt-ai/outlines
[4]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct#benchmarks
[5]https://fullfact.org/
[6]https://snopes.com/

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%	-	-	-	-

The PolitiFact label definitions, as specified in Section 3, are consistent across schemes. By evaluating these schemes, we aim to understand how different levels of granularity influence the model's ability to classify claims and provide useful explanations.

4.5. Evidence Retrieval

Although PolitiFact's fact-checking articles provide human-collected evidence that informs the justification and final verdict, extracting and decontextualizing these evidences is not trivial and requires additional specialized modeling and annotation. Consequently, in this study we focus on web-based fact-checking to gather relevant information. We collect the evidence by querying a web search API[7] for each claim and retrieve the top 10 search results. We do not apply any query optimization or re-ranking of results. We restrict the search to exclude a list of well-known US fact-checking sites as well as snippets that mention keywords such as "PolitiFact", "fact-check", or "debunk" to exclude fact-checking articles or direct references. This way, we aim to reduce information leaking in from pages reporting the actual verification results, rather than evidence. Due to these constrains, we were not able to retrieve evidences for 667 claims. Table 3 lists three search results for the claim presented in Figure 1 where we shortened the snippet text and removed the title and source URLs.

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 ...

4.6. Experimental Setup

To assess the performance of our automated fact-checking approach, we utilize a combination of classification and generation evaluation metrics. These metrics evaluate both the performance of verdict classification and the quality of generated outputs, ensuring a comprehensive analysis of system performance. We report accuracy and F1-Scores in different aggregations strategies to observe different aspects of the classification results. To evaluate the quality of generated outputs, we use TIGERScore, a reference-free metric that has been fine-tuned to assess generated text quality based on a set of criteria and assign penalties to mistakes [34]. Specifically, comprehension, accuracy, informativeness, and coherence are evaluated. TIGERScore provides an error evaluation of the generated outputs and assigns penalty scores between $[-5, -0.5]$ for each error without relying on ground truth references. The penalty scores are added up and reported for each case. Thus, a score close to 0 shows higher quality output. In this study, we utilize the 13B TIGERScore model with default hyperparameters to

[7] https://serper.dev/

evaluate generated outputs. The evaluation prompt design follows our task prompt as described in Section 4.2.
Due to the stochastic nature of LLMs, evaluation is often not trivial. Thus, we run each fact-checking task three times and report the majority vote for the classification performance evaluation. Additionally, as TIGERScore is a generative evaluation metric, we also run it three times and report the average metric for the justification quality assessment.

5. Evaluation

The evaluation section presents a detailed analysis of our automated fact-checking approach. We assess the task performance based on model size, labeling scheme, and the impact of evidence retrieval on both classification performance and the quality of generated outputs. This evaluation is structured around our predefined hypotheses and utilizes the previously introduced range of metrics to ensure a robust assessment. Additionally, statistical analyses are conducted to determine the significance of observed performance differences.

5.1. Hypotheses

Our evaluation focuses on several fundamental questions regarding the introduced problem setting. We examine whether models can reliably distinguish between the original truthfulness labels, or if alternative approaches to claim annotation and the fact-checking task formulation are required. We also consider potential limitations on the granularity of truthfulness labels that models can effectively handle. Additionally, we assess the role of parametric knowledge in task performance, specifically whether model size yields the expected effect of better performance. Finally, we investigate the impact of evidence integration on task performance. Based on these research questions, our evaluation is structured around the following hypotheses:

Hypothesis H_1: Classification task performance decreases as label complexity increases.
Hypothesis H_2: Justification quality decreases as label complexity increases.
Hypothesis H_3: Retrieving and incorporating evidence improves both classification accuracy and the quality of generated justifications.
Hypothesis H_4: Larger models perform better in the classification task and produce higher quality justifications.
Hypothesis H_5: Smaller models benefit more significantly from evidence integration than larger models due to less parametric knowledge being available.

5.2. Example Output

Previously, we introduced a claim involving the New York Times editorial and Sarah Palin in Table 1 and showcased examples of retrieved web evidence in Table 3. In Figure 2, we now present an actual output generated by the Llama3.3-70B model under the evidence-augmented setting with the five-class labeling scheme.
The output in Table 2 demonstrates good justification quality. The verdict is correctly classified as False, aligning with the evidence and reasoning provided. The reasoning section effectively incorporates the retrieved evidence, presenting a detailed analysis of the claim and referencing the correction issued by the New York Times. It also mentions the court's ruling in favor of the publication, which is not directly relevant to the claim verification. The explanation is concise and supports the verdict, accurately summarizing the key points without introducing ambiguity. This example highlights the potential of retrieval-augmented generation to improve classification accuracy and justification quality.

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, and the New York Times was sued 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.

5.3. Results

The results presented in Tables 4,5, and 6 illustrate the results in classification performance across different labeling schemes and model sizes, with and without evidence retrieval. For the five-class setup (Table 4), evidence retrieval consistently enhances model performance, as seen in higher $F1$ scores and TIGERScore improvements. However, the 3B model struggles to outperform the baseline significantly, indicating limited capacity in handling a complex task such as automated fact-checking.

Table 4
Results for 5 Classes.

Model	**Evidence**	$F1_{macro}$	$F1_{weighted}$	$F1_{micro}$	Acc.	TIGER↑
Baseline	-	0.2	0.213	0.213	0.213	-
3.2-3B-Instruct	No	0.216	0.242	0.273	0.273	-3.995
	Yes	0.259	0.285	0.321	0.321	-3.116
3.1-8B-Instruct	No	0.244	0.268	0.293	0.293	-3.416
	Yes	0.278	0.301	0.339	0.339	-2.578
3.1-70B-Instruct	No	0.314	0.338	0.356	0.356	-2.554
	Yes	0.335	0.359	0.389	0.389	-2.466
3.3-70B-Instruct	No	0.314	0.337	0.357	0.357	-2.361
	Yes	0.351	0.375	**0.405**	0.405	**-1.686**

In the three-class classification scheme (Table 5), evidence retrieval again provides a notable performance boost across all models, with improvements becoming more pronounced in larger models. This indicates that as label complexity decreases, models are better able to leverage evidence to enhance classification accuracy and justifications. The 3.3-70B-Instruct model achieves the highest scores, emphasizing the advantage of size when combined with external knowledge. Table 5 presents the classification metrics for the three-class scheme. Similar to the five-class results, evidence retrieval enhances performance across all models.

For binary classification results in Table 6, the reduced complexity of the task yields the highest overall performance across all models. Evidence retrieval continues to provide a measurable benefit, particularly in the largest models, where the highest $F1$ scores and TIGERScore improvements are observed.

In the following, we investigate the indications we have described earlier with statistical analyses to draw conclusions about the hypotheses we specified in Section 5.1.

We conducted a Friedman test on $F1_{micro}$ across the three classification schemes, with and without evidence. The result indicates that at least one of the schemes differs significantly ($p < 0.05$) in terms of classification performance. These findings support hypothesis H_1 that as labeling becomes more complex, classification performance tends to decrease, potentially due to more nuanced distinctions between labels that increase the quantity of prediction errors.

Table 5
Results for 3 Classes.

Model	**Evidence**	$F1_{macro}$	$F1_{weighted}$	$F1_{micro}$	Acc.	TIGER↑
Baseline		0.333	0.371	0.371	0.371	-
3.2-3B-Instruct	No	0.322	0.396	0.464	0.464	-4.069
	Yes	0.390	0.453	0.498	0.498	-3.205
3.1-8B-Instruct	No	0.389	0.443	0.472	0.472	-3.391
	Yes	0.448	0.506	0.525	0.525	-2.751
3.1-70B-Instruct	No	0.499	0.551	0.542	0.542	-2.610
	Yes	0.521	0.571	0.556	0.556	-2.524
3.3-70B-Instruct	No	0.524	0.570	0.556	0.556	-2.383
	Yes	0.550	0.601	**0.589**	0.589	**-1.884**

Table 6
Results for Binary Classification.

Model	**Evidence**	$F1_{macro}$	$F1_{weighted}$	$F1_{micro}$	Acc.	TIGER↑
Baseline	-	0.500	0.501	0.501	0.501	-
3.2-3B-Instruct	No	0.624	0.624	0.624	0.504	-3.870
	Yes	0.647	0.647	0.647	0.557	-3.150
3.1-8B-Instruct	No	0.649	0.649	0.649	0.543	-3.367
	Yes	0.668	0.668	0.668	0.589	-2.741
3.1-70B-Instruct	No	0.689	0.689	0.689	0.684	-2.560
	Yes	0.708	0.708	0.708	0.691	-2.433
3.3-70B-Instruct	No	0.722	0.722	0.722	0.707	-2.303
	Yes	0.747	0.747	**0.747**	0.734	**-1.739**

For TIGERScore evaluation, the Friedman test was significant ($p < 0.05$) for the setting with evidences, but a subsequent Conover's test revealed no significant pairwise differences. Additionally, the Friedman test reveals no significance for the setting without evidence. This result suggests that there is no measurable difference in justification quality across the three schemes, with or without evidence. These findings reject hypothesis H_2 that more complex label sets negatively affect overall justification quality. This may be in part due to claim analysis and explanation being difficult enough, regardless of whether the label schemes are more or less complex.

To determine the statistical significance of performance when including evidence, we conducted paired t-tests comparing models with and without evidence across all classification schemes for the $F1_{micro}$ and TIGERScore. The results indicate a statically significant difference (p < 0.01) for both metrics when evidence retrieval is included during the fact-checking task. This supports hypothesis H_3. Thus, external evidence helps the model to disambiguate classes and produce more useful justifications for the fact-checking task.

To evaluate whether larger models outperform their smaller counterparts, we performed a Friedman test on both $F1_{micro}$ and TIGERScore with and without evidence across four different model sizes. The results indicate a significant difference ($p < 0.05$), confirming that model size has a measurable impact on performance. These findings support hypothesis H_4.

Finally, to investigate whether smaller models benefit more from evidence integration than larger models, we examined the performance gains by subtracting the no evidence scores from the with evidence scores for both $F1_{micro}$ and TIGERScore across all model sizes. For $F1_{micro}$ gains, the Friedman test showed no significant difference ($p = 0.167$), whereas the TIGERScore gains were statistically significant ($p < 0.05$). Thus, we partially reject hypothesis H_5. This implies that larger models benefit even more from external evidence, presumably due to their ability to reason effectively across long context sizes, whereas smaller models exhibit relatively limited improvements. We expect that integrating more credible and complete information sources, could enhance overall performance for both smaller and larger models even further.

5.4. Ablation Study

Since we consider fine-tuning approaches impractical for real-world automated fact-checking, due to the dynamic and fast-changing nature of misinformation, which limits the usefulness of models trained on static datasets, our primary focus in this study has been on few-shot inference using large language models. Earlier encoder-based architectures, such as BERT, were constrained by a maximum sequence length of 512 tokens, which restricted their ability to incorporate additional context. Recent advancements, such as ModernBERT [35], adjust the original BERT architecture and are able to support sequence lengths of up to 8192 tokens. This allows the integration of more contextual information and retrieved evidence directly into the classification process, enabling the evaluation of their utility for veracity prediction.

To complement our few-shot evaluation and to better understand how different input signals contribute to classification outcomes, we conduct an ablation study using the ModernBERT-large architecture. Specifically, we fine-tune the model across a series of input configurations to assess how predictive performance changes when incrementally adding additional contextual information. We begin with the claim alone as input. We then add information about the surrounding context in which the claim appeared, such as a speech, interview, or social media post. Next, we incorporate the speaker who issued the claim. Finally, we include retrieved web evidence that provides external factual grounding. This study helps quantify the individual impact of each component and provides an empirical upper predict performance bound for fine-tuning on the dataset, enabling a more informed comparison with few-shot LLM performance.

Table 7
Results for classification fine-tuning ModernBERT-large on different configurations reported on $F1_{micro}$.

Model	**Input Configuration**	**Five-Class**	**Three-Class**	**Binary**
ModernBERT	Claim	0.239	0.363	0.637
	Claim + Context	0.276	0.375	0.612
	Claim + Context + Speaker	0.287	0.445	0.624
	Claim + Context + Speaker + Web	0.317	0.491	0.696
3.3-70B-Instruct	Claim + Context + Speaker + Web	0.357	0.578	0.721

The results in Table 7 show that incorporating evidence consistently produces the most significant gains across all label granularities. In the five-class setting, starting with only the claim results in the lowest performance. Adding context leads to modest improvements, suggesting that surrounding details help disambiguate some claims. For example, knowing whether a statement was made during a campaign rally or in an official policy document can influence its interpretation. Speaker information further improves performance, which may be attributed to prior knowledge about the speaker's reliability, role, or political alignment that implicitly guides veracity estimation. In the binary setting, adding context does not improve performance and even reduces it slightly. This outcome likely stems from the way binary labels are constructed by merging more nuanced classes. As a result, different claims with dissimilar contexts may be grouped under the same binary label, making context a noisy feature. In contrast, speaker information helps more consistently. This may reflect the fact that in coarse-grained tasks, speaker identity acts as a high-level signal about the probable factuality of a claim.

The classification results align closely with the previously presented LLM few-shot inference results, showing that evidence consistently provides significant performance improvements across all label schemes. For both classifiers and LLMs, the inclusion of evidence enables better disambiguation and enhances predictive performance, particularly in more complex multi-class tasks. While smaller LLMs provide comparable performance across tasks in few-shot inference, larger LLMs consistently surpass the fine-tuned SLMs without requiring fine-tuning and provide the additional advantage of generating reasoning and detailed justifications across more extensive contexts. This highlights the general utility of LLMs for AFC.

6. Discussion and Conclusion

This study investigated AFC of real-world claims using LLMs in a few-shot inference scenario. By evaluating task performance across three labeling schemes and multiple LLM sizes of the same architecture, we demonstrated the importance of evidence integration, model scale, and labeling complexity in determining system effectiveness. Evidence retrieval consistently improved classification accuracy and justification quality, with larger models showing the most significant gains. In contrast, smaller models struggled to perform or benefit as much from evidence integration, highlighting the need for further optimization in computationally constrained environments. While more coarse-grained labels naturally yield higher performance, future work should explore how to integrate alternative labeling strategies and a more nuanced assessment of claims across different perspectives to develop a robust AFC approach. Our experiments show that LLMs can effectively perform multi-component tasks by reasoning over presented data and generating detailed justifications. However, our results also indicate that alternative approaches leveraging fine-tuning can be advantageous for specific subtasks or in resource-constrained settings. For instance, while LLMs excel in knowledge-based reasoning and explanation generation in few-shot scenarios, models like ModernBERT can be sufficiently effective for classification tasks when supervised training data is available. This suggests the potential for hybrid frameworks where supervised fine-tuning is employed for tasks that rarely change, such as document type or natural language inference classification, while LLMs are reserved for dynamic scenarios that require integration of dynamic facts to produce grounded reports. Moreover, integrating more credible evidence, including human-aggregated sources, could further enhance AFC performance by providing more reliable context for claim evaluation. Furthermore, although our study focused on LLMs from the Llama family, future work could benefit from expanding the comparison to include different model families. A broader analysis comparing diverse model families would be valuable, especially for applications where training and inference costs, use cases, and interpretability requirements differ substantially. To extend AFC toward intelligent decision assistance for expert fact-checkers, future research should focus on structuring justifications to align more closely with human verification strategies. This includes presenting concise, faithful explanations that detail key reasoning steps and clearly highlight the integrated evidence. Preliminary observations indicate that LLM-based systems can suffer from hallucinations, underscoring the need for extensive evaluation and user studies to understand how experts interpret and trust the generated explanations. Such studies would not only help refine the presentation of justifications but also identify gaps in current AFC systems and better define their role in supporting, rather than replacing, human fact-checking efforts.

Acknowledgments

This study is partially funded by the German Federal Ministry of Education and Research (BMBF, reference: 03RU2U151C) in the scope of the research project news-polygraph and by JST AIP Acceleration Research (JPMJCR24U3) and JST CREST Grants (JPMJCR20D3).

References

[1] S. Lewandowsky, J. Cook, D. Lombardi, Debunking Handbook 2020, 2020. doi:10.17910/B7.1182.
[2] A. Vlachos, S. Riedel, Fact Checking: Task definition and dataset construction, in: C. Danescu-Niculescu-Mizil, J. Eisenstein, K. McKeown, N. A. Smith (Eds.), Proceedings of the ACL 2014

Workshop on Language Technologies and Computational Social Science, Association for Computational Linguistics, Baltimore, MD, USA, 2014, pp. 18–22. doi:10.3115/v1/W14-2508.

[3] N. Hassan, F. Arslan, C. Li, M. Tremayne, Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Halifax NS Canada, 2017, pp. 1803–1812. doi:10.1145/3097983.3098131.

[4] Z. Guo, M. Schlichtkrull, A. Vlachos, A Survey on Automated Fact-Checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206. doi:10.1162/tacl_a_00454.

[5] B. Nyhan, Facts and Myths about Misperceptions, Journal of Economic Perspectives 34 (2020) 220–236. doi:10.1257/jep.34.3.220.

[6] S. Cresci, A. Trujillo, T. Fagni, Personalized Interventions for Online Moderation, in: Proceedings of the 33rd ACM Conference on Hypertext and Social Media, HT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 248–251. doi:10.1145/3511095.3536369.

[7] B. Thompson, How to battle misinformation with Sander van der Linden, Nature (2023). doi:10.1038/d41586-023-00899-0.

[8] X. Zhou, R. Zafarani, A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities, ACM Computing Surveys 53 (2021) 1–40. doi:10.1145/3395046. arXiv:1812.00315.

[9] L. Graves, Understanding the promise and limits of automated fact-checking, 2018.

[10] P. B. Brandtzaeg, A. Følstad, Trust and distrust in online fact-checking services, Communications of the ACM 60 (2017) 65–71. doi:10.1145/3122803.

[11] V. Setty, Surprising Efficacy of Fine-Tuned Transformers for Fact-Checking over Larger Language Models, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Washington DC USA, 2024, pp. 2842–2846. doi:10.1145/3626772.3661361.

[12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. doi:10.48550/arXiv.1810.04805. arXiv:1810.04805.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training (2018).

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2021. doi:10.48550/arXiv.2005.11401. arXiv:2005.11401.

[15] W. Ferreira, A. Vlachos, Emergent: A novel data-set for stance classification, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 1163–1168. doi:10.18653/v1/N16-1138.

[16] N. Kotonya, F. Toni, Explainable Automated Fact-Checking: A Survey, 2020.

[17] D. Russo, S. S. Tekiroğlu, M. Guerini, Benchmarking the Generation of Fact Checking Explanations, Transactions of the Association for Computational Linguistics 11 (2023) 1250–1264.

[18] I. Eldifrawi, S. Wang, A. Trabelsi, Automated Justification Production for Claim Veracity in Fact Checking: A Survey on Architectures and Approaches, 2024. arXiv:2407.12853.

[19] L. Pan, X. Wu, X. Lu, A. T. Luu, W. Y. Wang, M.-Y. Kan, P. Nakov, Fact-Checking Complex Claims with Program-Guided Reasoning, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6981–7004.

[20] H. Wang, K. Shu, Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6288–6304. doi:10.18653/v1/2023.findings-emnlp.416.

[21] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On Faithfulness and Factuality in Abstractive Summarization, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational

Linguistics, Online, 2020, pp. 1906–1919. doi:10.18653/v1/2020.acl-main.173.
[22] C. Si, N. Goyal, S. T. Wu, C. Zhao, S. Feng, H. Daumé III, J. Boyd-Graber, Large Language Models Help Humans Verify Truthfulness – Except When They Are Convincingly Wrong, 2024.
[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023.
[24] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. G. Simonsen, MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims, in: Proceedings of 2019 EMNLP-IJCNLP, Association for Computational Linguistics, 2019, pp. 4685–4697.
[25] W. Y. Wang, "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. doi:10.18653/v1/P17-2067.
[26] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 2733–2743. doi:10.1145/3539618.3591879. arXiv:2205.12487.
[27] N. Kotonya, F. Toni, Towards a Framework for Evaluating Explanations in Automated Fact Verification, 2024. arXiv:2403.20322.
[28] Z. Zhang, K. Rudra, A. Anand, Explain and Predict, and then Predict Again, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 418–426. doi:10.1145/3437963.3441758.
[29] P. Atanasova, J. G. Simonsen, C. Lioma, I. Augenstein, Generating Fact Checking Explanations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 7352–7364.
[30] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. Da Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, P. Resnik, The Prompt Report: A Systematic Survey of Prompting Techniques, 2024. arXiv:2406.06608.
[31] B. T. Willard, R. Louf, Efficient Guided Generation for Large Language Models, 2023.
[32] A. Hanselowski, C. Stab, C. Schulz, Z. Li, I. Gurevych, A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking, 2019. doi:10.48550/arXiv.1911.01214. arXiv:1911.01214.
[33] M. Schlichtkrull, Z. Guo, A. Vlachos, AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web, 2023. doi:10.48550/arXiv.2305.13117. arXiv:2305.13117.
[34] D. Jiang, Y. Li, G. Zhang, W. Huang, B. Y. Lin, W. Chen, TIGERScore: Towards Building Explainable Metric for All Text Generation Tasks, 2024.
[35] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, 2024. doi:10.48550/arXiv.2412.13663. arXiv:2412.13663.