Checklist-Prompted Feature Extraction for Interpretable and Robust Claim Check Worthiness Prediction

Yuka Teramoto[1], **Takahiro Komamizu**[2], **Mitsunori Matsushita**[3], **Kenji Hatano**[1]

[1]Doshisha University, 1–3 Tatara-Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
[2]Nagoya University, Furo, Chikusa, Nagoya, Aichi 464-8603, Japan
[3]Kansai University, 2–1–1 Ryozenji, Takatsuki, Osaka 569-1095, Japan
{teramoto, hatano}@mil.doshisha.ac.jp, taka-coma@acm.org, m_mat@kansai-u.ac.jp

Abstract

This study explores the use of Large Language Models (LLMs) for Claim Check Worthiness Prediction (CCWP), a critical first step in fact-checking. Building on prior research, we propose a method that utilizes structured checklists to break down CCWP into interpretable and manageable subtasks. In our approach, an LLM answers 52 human-crafted questions for each claim, and the resulting responses are used as features for traditional supervised learning models. Experiments across six datasets show that our method consistently improves performance across key evaluation metrics, including accuracy, F_1 score, surpassing few-shot prompting baselines on most datasets. Moreover, our method enhances the stability of LLM outputs, reducing sensitivity to prompt design. These findings suggest that LLM-based feature extraction guided by structured checklists offers a promising direction for more reliable and efficient claim prioritization in fact-checking systems.You can access and utilize the program and code at the following GitHub repository: [1]

Introduction

Fact-checking is an effective countermeasure against the spread of misinformation when conducted rapidly and with sufficient investment in careful verification. Misinformation poses a growing threat in today's information society: it can disrupt democratic processes by damaging the reputations of election candidates, harm public health through false claims about diseases like COVID-19, and foster social isolation by promoting conspiracy theories (Das et al. 2023; Schmitt et al. 2024).

It is necessary to automate fact-checking the ever-increasing volume of misinformation efficiently. Rapid completion of fact-checking can more effectively curb the spread of harmful misinformation (Rastogi and Bansal 2023). However, the potentially misleading claims often exceed the processing capacity of fact-checkers. Fact-checking is a complex and time-consuming process that may take days or even weeks, placing a significant burden on individuals. These facts highlight the importance of computational tools in fact-checking mechanisms, as discussed in previous research (Guo, Schlichtkrull, and Vlachos 2022).

[1]https://github.com/TRMT-Yuka/FactCheck/CPFE_in_MisD25

We explore using LLM, recently gaining significant attention, to support parts of the fact-checking workflow. LLMs have demonstrated strong performance across various tasks, prompting a paradigm shift. However, many studies caution against relying on LLM for fact-checking, as they are prone to biases and hallucinations[2], and may amplify inaccurate information (Neumann et al. 2024). Instead, prior work suggests that human–LLM collaboration is a more practical approach to fact-checking (Das et al. 2023; Schmitt et al. 2024), with humans intervening in critical stages to mitigate risks.

In this context, one promising use case for LLM is the automation of claim prioritization for verification (Majer and Šnajder 2024). This task, known as **Claim Check-Worthiness Prediction (CCWP)**, is a crucial first step in the fact-checking workflow. In today's information society, web users generate an overwhelming number of posts, and each claim differs in its potential social impact. Therefore, fact-checkers must screen these claims to determine whether they should enter the workflow.

In this study, we investigate the role of LLM in a human-in-the-loop fact-checking framework by introducing the structured checklist designed explicitly for CCWP. A key barrier to automating CCWP lies in the ambiguity of value judgment criteria, a challenge noted in previous studies (Nenno 2024; Neumann and Wolczynski 2023; Gencheva et al. 2017a). Identifying fine-grained factors influencing check-worthiness and structuring the assessment process is essential for fair and effective claim prioritization and automation.

The structured checklist we adopt (Sehat et al. 2024), developed through interviews with 23 professional fact-checkers, assesses the potential harms of misinformation across five dimensions: Fragmentation, Actionability, Believability, Likelihood of spread, and Exploitativeness. We investigate whether LLM can leverage this checklist to help identify which claims fact-checkers should prioritize for urgent verification. This checklist-based approach offers several advantages: First, it decomposes the complex CCWP task into smaller, more interpretable, and independent sub-

[2]"Can ChatGPT fact-check? PolitiFact tested." https://www.politifact.com/article/2023/may/30/can-chatgpt-fact-check-politifact-tested/, Accessed: March 29, 2025.

Table 1: Examples of questions from the structured checklist which used to determine check-worthiness

Question Category	Example
External information	Is there a lack of high-quality information that is publicly accessible and refuting the message's claim?
Impressions from the text	Does the message directly call audience members to share the content further?

tasks. Second, it reduces task complexity and facilitates human oversight. Third, it yields more robust performance than directly prompting LLM for CCWP.

The contributions of this study are as follows:

- We propose a novel approach for CCWP that decomposes the task into a set of fine-grained sub-decisions based on the structured checklist.
- We demonstrate that using LLM-generated responses to these sub-decisions as features for supervised classifiers improves predictive performance over few-shot prompting baselines.
- Our method provides more excellent stability than LLM prompting, mitigating variability due to prompt design and model differences.
- We conduct extensive experiments across six publicly available datasets and provide a unified model performance evaluation using both classification and ranking metrics.

Related Work

Many previous studies have addressed the task of CCWP (Hassan, Li, and Tremayne 2015; Patwari, Goldwasser, and Bagchi 2017; Gencheva et al. 2017b; Gangi Reddy et al. 2022; Majer and Šnajder 2024). The CLEF 2024 CheckThat! Lab provides both datasets and a competitive platform for check-worthiness tasks(Barrón-Cedeño et al. 2024). More recently, researchers have expanded their focus to using LLMs in subtasks of fact-checking workflow, including CCWP (Quelle and Bovet 2024).

As mentioned in the Introduction Section, LLMs should not be used to perform CCWP directly because LLMs contain biases and can amplify these biases (Neumann et al. 2024). Moreover, LLMs can produce hallucinations that deviate from facts, and the mechanisms of these hallucinations are not entirely understood (Ji et al. 2023). In an evaluation using LLMs to perform the task directly, an F_1 score of 0.75 was achieved on the ClaimBuster dataset, and relatively high evaluation metrics were reported even with simple zero-shot and few-shot prompts (Majer and Šnajder 2024). In this study, we use these results as a baseline.

The CCWP task is crucial for improving the efficiency of fact-checking; however, ambiguity in the criteria for check-worthiness remains a significant bottleneck in automating fact-checking value assessments (Nenno 2024). This ambiguity poses a challenge for the automation of CCWP in computer science and fairness-related research in the social sciences. Indeed, factors such as the race, beliefs, expertise, and minority status of fact-checkers often affect these judgments, raising concerns about equity and bias in fact-checking systems (Neumann and Wolczynski 2023; Gencheva et al. 2017a).

Some Researchers in the social sciences domain aim to explain variations in fact-checkers' judgments and to reduce ambiguity to objectively analyze the validity of their decisions. One promising direction in this line of work is to divide the fact-checking process into multiple, more fine-grained decision-making tasks, especially those involving value-laden assessments. Sehat et al. created the structured checklistto determine the priority of fact-checking (Sehat et al. 2024). This checklist was developed based on a fact-checking survey of human experts. Examples of the questions are shown in Table 1. Fact-checkers respond to each item with "yes," "no," or "unknown." The higher the number of "yes" responses, the more urgent it is to fact-check the corresponding claim. The core idea of this research is to apply the structured checklistas a prompt for LLM.

Proposed Method

Figure 1 shows how the features were generated in our experiment. We use the structured checklist (Sehat et al. 2024) outlining human fact-checkers' steps to prioritize claims.

In this paper, the task of *Claim Check-Worthiness Prediction* for fact-checking is modeled as a binary classification problem that a predictor indicates the check-worthiness of the claim. The goal is to identify whether a given claim is worth being fact-checked. Formally, let C denote a claim, and the objective is to learn a function: $f : C \rightarrow \{0, 1\}$ such that:

$$f(C) = \begin{cases} 1 & \text{if } C \text{ is check-worthy} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

As the previous research indicates a promising result (Teramoto et al. 2024), this paper assumes that LLMs will likely perform comparably to humans on some of these questions in the structured checklist (Teramoto et al. 2024). Indeed, even in behavioral economics, it is known that LLMs can mimic human impressions and value judgments (Leng 2024; Wang et al. 2023). Our preliminary research also showed similar results.

We observed the structured checklist and hypothesized that two types of these questions exist: those that ask for external information relevant to the claim and those that ask for characteristics or impressions derived from the text. The former includes items such as the characteristics of the claim's

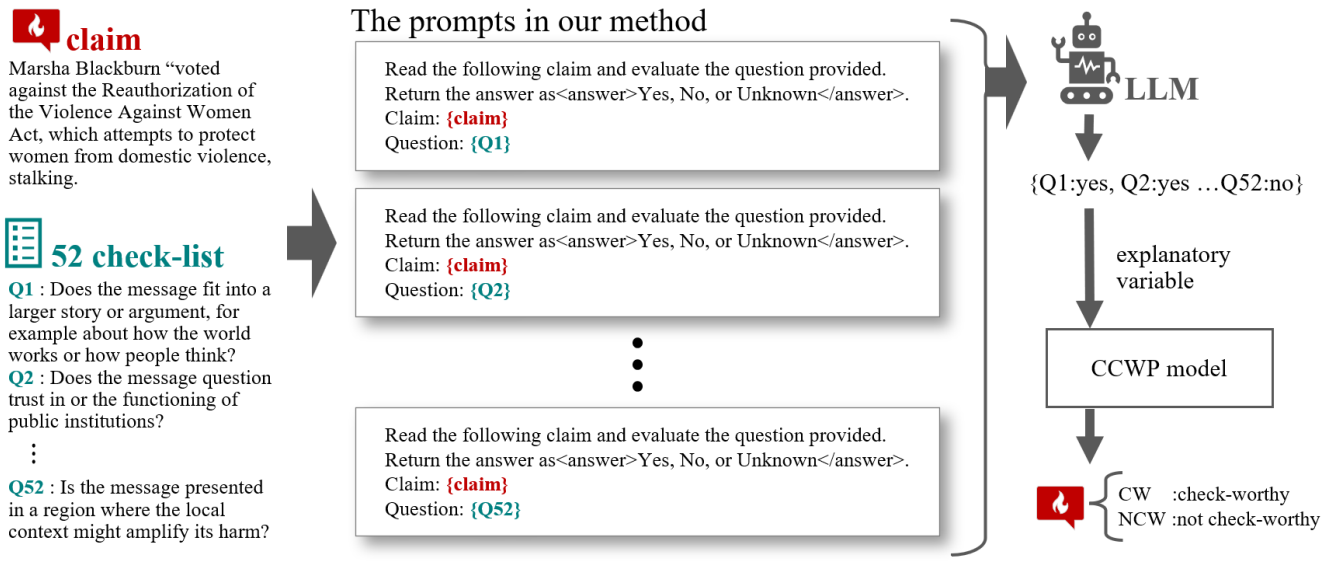

Figure 1: Our proposed method

issuer and whether there have been any official announcements related to the claim. The latter includes whether the claim makes statements about global trends or contains aggressive bias against specific groups. The latter group of questions pertains to human impressions and value judgments derived from the text, which LLMs can likely replicate sufficiently. On the other hand, the former questions might be influenced by external factors. Examples of such external factors include the context of the documents accessible to the LLM and the structure of the websites where the claims are found. Therefore, there may be differences in the accuracy of LLM responses. The possible existence of these two groups indicates that the questions in the structured checklist are not equally important to predict check-worthiness of a claim.

Based on this idea, this paper proposes a data-oriented CCWP model that leverages the answers to these questions in the structured checklist by LLMs. Let C denotes a claim, L denotes a set of questions in the structured checklist, and $a_\ell = \text{LLM}(C, \ell)$ denotes an answer to a question $\ell \in L$ for C by an LLM. Then, the predictor f receives the set of answers $A = \{a_\ell \mid \ell \in L\}$ to C and predicts the check-worthiness of C as defined in Equation 1.

Figure 2 shows the prompt used for obtaining answers to questions in the structured checklist. {claim} and {question} are placeholders for claim C and question $\ell \in L$. Following the precedent set by prior research using LLMs as annotators (Leng 2024), the response sections are structured with tags. The sentences following "Claim:" and "Question:" will be modified. For reviewing the claim, it is necessary to answer the 52 questions created in prior research (Sehat et al. 2024). Therefore, we will create 52 different prompts for each claim.

> Read the following claim and evaluate the question provided. Return the answer as <answer>Yes, No, or Unknown</answer>.
> Claim: {claim}
> Question: {question}

Figure 2: Prompt to LLM: {claim} and {question} are placeholders for claim and question, and LLMs are asked to answer the question in Yes, No, or Unknown with tags.

Experimental Evaluation

Task Settings

Dataset In the CCWP task field, other major datasets exist, including CLEF CheckThat! Lab 2019, 2021 (Elsayed et al. 2019; Nakov et al. 2021), TATHYA (Patwari, Goldwasser, and Bagchi 2017), ClaimRank (Jaradat et al. 2018), and PoliClaim (Gencheva et al. 2017b). These datasets are primarily situated in the debate domain, where the task is to assess the check-worthinessof individual sentences extracted from political debates. However, recent research has pointed out that check-worthinessassessment in this domain often lacks surrounding context, highlighting the need for new methods to supplement missing contextual information. Since the present study does not address the challenge of contextual supplementation, we exclude all debate-domain datasets except for ClaimBuster, which traditionally used important data sets in comparative experiments.

Moreover, items from the structured checklist adopted in this study are specifically designed to identify factors contributing to check-worthiness. They do not target detecting whether a sentence contains a verifiable claim. We exclude

the NewsClaims dataset (Gangi Reddy et al. 2022) in our study, as it does not explicitly specify whether it is designed for the CCWP task. It is also well known that the linguistic distribution in training data can affect the performance of LLMs. However, we do not address this issue, so all experiments in this study are conducted using English-only data. Accordingly, we also exclude IndianClaims (Jha et al. 2023) from our evaluation.

On the other hand, NL4IF (Shaar et al. 2021) satisfies all of the experimental conditions defined in our study and is newly included despite not being referenced in prior work. This dataset contains tweets related to COVID-19, annotated across multiple fact-checking dimensions, including verifiability, falsity, public interest, harmfulness, need for verification, social impact, and government intervention. Among these, we only utilize the subset of the dataset explicitly labeled for "need of verification" in this study.

Evaluation Metrics We employ standard classification metrics to evaluate the performance of each method. Specifically, we report macro-averaged $Precision$ (P), $Recall$ (R), and F_1 score (F_1-M), as well as the micro-averaged F_1 score (F_1-μ). In addition, we include class-wise F_1 scores, denoted as F_1*class label name*, to provide a more fine-grained performance analysis. Among the reported metrics, the best-performing scores for each evaluation criterion are highlighted in bold.

Settings of Proposed Method

LLMs in the Proposed Method We adopted Llama 3 with a parameter size of 8B for the proposed method and the follow-up experiments of previous studies. In our method, the number of queries to LLM is enormous, at 52 times per claim for which the check value is determined. For this reason, it is not realistic to use paid APIs like GPT-4 (Achiam et al. 2023) in situations where the proposed method is actually used. For this reason, we adopted a relatively lightweight model that can run locally. In order to eliminate randomness as much as possible and obtain stable answers in tasks related to fact-checking, the value of the temperature parameter is set to 0.

Predictor Models in the Proposed Method

In this study, we used typical supervised machine learning methods as a model that uses the structured checklist judgment results by the LLM as the explanatory variable on the CCWP task. We selected the following models to represent a diverse set of learning paradigms.

Logistic Regression (LR) : This linear classification method uses the logistic function to model the probability of categorical outcomes. It is appreciated for its simplicity and ease of interpretation.

Decision Tree (DT) : This non-linear model divides data using hierarchical, feature-based splits, providing a highly interpretable and intuitive structure.

Random Forest (RF) : This method combines multiple decision trees trained on different bootstrap samples, and by aggregating their predictions, it achieves greater robustness and mitigates overfitting.

Gradient Boosting (GB) : This ensemble method builds weak learners sequentially to correct previous errors, leading to high accuracy on structured data.

Neural Network (NN) : This layered model learns complex, non-linear relationships by applying linear transformations followed by activation functions, making them suitable for a wide range of prediction tasks.

Baseline As a comparison with the proposed method, a previous study (Majer and Šnajder 2024) used a small number of LLM prompts to perform the CCWP task. We will conduct a follow-up experiment. In the previous study, only F_1 score was used as an evaluation metric. In this study, however, it is necessary to compare the performance of each label in detail, so that we can conduct a follow-up experiment. This is because the most important is the ability to identify the assertion to be verified accurately, that is, the labeled data of class CW. Since many datasets overlap, the same prompts as in the previous study will be used in the follow-up experiment. Of the datasets not used in the previous study, those that share the same prompts as the CLEF CheckThat! Lab series will be used. In addition, for the NLP4IF dataset not used in the previous study, these prompts will be used because it has the same domain and structure as the CLEF CheckThat! Lab series. In order to ensure that the data is collected, we added a prompt to the previous study to enclose the response in the `<answer></answer>` tag.

In previous research, there were several categories with different granularities of context included in the prompts. Therefore, we adopted the prompts referred to as V2-type as representative for the follow-up experiment. This is because it was reported that the V2-type prompts consistently provided the best or second best performance across a wide range of datasets.

Experimental Results and Discussion

The experimental results are presented in Table 3a. The column labeled LLM corresponds to the baseline performance obtained via few-shot prompting. Our proposed method yielded substantial improvements in Acc ($accuracy$), macro-averaged F_1, and micro-averaged F_1 across four of the six datasets, with the exception of CT21 and CT22. In the context of the CCWP task, these are used as the primary evaluation metrics that collectively capture overall classification performance. Notable improvements include a 43-point gain in Acc, a 42-point gain in macro-averaged F_1 on the CB dataset, and a 42-point improvement in micro-averaged F_1 on the NLP4IF dataset.

On the other hand, when focusing solely on the CW label—which we place the most importance on—the F_1 score of the baseline (few-shot prompting) exceeded that of our proposed method across all datasets.

The accuracy of our follow-up experiment using the CB dataset was notably lower than that reported in the previous study. This degradation is likely attributable to the additional prompt instructions introduced during data preprocessing, which required the model to enclose its output within specific tags(`<answer></answer>`). In LLMs with a rela-

Table 2: Overview of the datasets used in our experiments. The column Label uses the abbreviations CFS (check-worthy factual statement), UFS (unimportant factual statement), NFS (non-factual statement), CW (check-worthy), and NCW (not check-worthy).

Abbr.	Dataset	Domain	Label
CB	ClaimBuster	debates	CFS UFS NFS
CT20	CLEF CheckThat! Lab 2020 Task1	X(Twitter)	CW NCW
CT21	CLEF CheckThat! Lab 2021 Task1A	X(Twitter)	CW NCW
CT22	CLEF CheckThat! Lab 2022 Task1A	X(Twitter)	CW NCW
NLP4IF	NLP4IF	X(Twitter)	CW NCW
ENV	Environmental Claims	reports	CW NCW

tively small number of parameters, such tag-based outputs tended to be unstable. In datasets excluding CB, the labels used in the prompting were binary (yes/no) and shared across multiple tasks and question items. In contrast, the CB dataset employed task-specific labels—CFS, UFS, and NFS—which may have increased task complexity and contributed to unstable model behavior.

As a new conclusion based on the above, one of the key strengths of our proposed method is its robustness in response generation. In few-shot prompting with LLMs, the final output is highly sensitive to the quality and structure of the prompt design. Our method, by comparison, allows for stable extraction of feature values when the question items and prompt format are fixed, and its performance remains consistent even across different datasets and model types. To support this observation, we conducted an additional experiment involving five-fold cross-validation on the training data, and the variances in accuracy and F_1 scores in each class label are summarized in Table 4.

Conclusion and Future Challenges

In this study, we proposed a novel method for Claim Check-Worthiness Prediction (CCWP) using the structured checklist to guide LLM-based feature extraction. Unlike few-shot prompting baselines, which often suffer from unstable outputs and sensitivity to prompt design, our method provides a more robust and interpretable approach by decomposing complex judgments into modular components. Experimental results across six datasets showed substantial improvements in core evaluation metrics—including accuracy, macro-averaged F_1, and micro-averaged F_1 on four datasets, while revealing limitations in predicting check-worthiness.

Table 3: Experimental results for CCWP models.

(a) CB

Metrics	baseline	LR	DT	RF	GB	NN
P	**0.42**	0.23	0.23	0.23	0.23	0.23
R	**0.34**	0.33	0.33	0.33	0.33	0.33
$F1$-μ	0.26	**0.70**	**0.70**	0.69	**0.70**	**0.70**
$F1$-M	0.15	**0.27**	**0.27**	**0.27**	**0.27**	**0.27**
$F1$-CFS	0.00	0.00	0.00	0.00	0.00	0.00
$F1$-UFS	**0.40**	0.00	0.00	0.00	0.00	0.00
$F1$-NFS	0.04	**0.82**	**0.82**	**0.82**	**0.82**	**0.82**

(b) CT20

Metrics	baseline	LR	DT	RF	GB	NN
P	**0.73**	0.58	0.58	0.60	0.60	0.57
R	**0.69**	0.55	0.57	0.59	0.57	0.56
$F1$-μ	**0.65**	0.59	0.59	0.61	0.61	0.59
$F1$-M	**0.64**	0.53	0.57	0.58	0.56	0.56
$F1$-CW	**0.70**	0.36	0.46	0.47	0.41	0.44
$F1$-NCW	0.59	0.70	0.67	0.70	**0.71**	0.67

(c) CT21

Metrics	baseline	LR	DT	RF	GB	NN
P	**0.54**	0.50	0.50	0.49	0.49	0.51
R	**0.68**	0.51	0.49	0.47	0.45	0.55
$F1$-μ	0.45	0.51	0.64	0.65	0.67	**0.71**
$F1$-M	0.37	**0.51**	0.43	0.43	0.43	0.48
$F1$-CW	0.16	**0.51**	0.09	0.08	0.06	0.12
$F1$-NCW	0.59	0.51	0.78	0.78	0.80	**0.83**

(d) CT22

Metrics	baseline	LR	DT	RF	GB	NN
P	0.63	**0.66**	0.59	0.59	0.62	0.58
R	**0.65**	0.52	0.53	0.52	0.53	0.54
$F1$-μ	0.49	**0.78**	0.77	0.77	0.77	0.75
$F1$-M	0.49	0.49	0.51	0.50	0.51	**0.54**
$F1$-CW	**0.46**	0.11	0.15	0.13	0.14	0.23
$F1$-NCW	0.52	**0.87**	0.86	**0.87**	**0.87**	0.85

(e) ENV

Metrics	baseline	LR	DT	RF	GB	NN
P	0.45	0.66	0.68	0.68	0.68	**0.74**
R	0.46	0.52	0.52	0.52	0.52	**0.53**
$F1$-μ	0.54	0.75	0.75	0.75	0.75	**0.76**
$F1$-M	0.36	0.48	0.47	0.47	0.47	**0.50**
$F1$-CW	**0.52**	0.11	0.08	0.08	0.08	0.14
$F1$-NCW	0.55	**0.86**	**0.86**	**0.86**	**0.86**	**0.86**

(f) NLP4IF

Metrics	baseline	LR	DT	RF	GB	NN
P	0.59	0.57	0.55	0.58	**0.74**	0.58
R	**0.67**	0.53	0.53	0.54	0.53	0.55
$F1$-μ	0.45	0.85	0.83	0.84	**0.87**	0.84
$F1$-M	0.43	0.53	0.53	0.54	0.53	**0.55**
$F1$-CW	**0.32**	0.15	0.16	0.17	0.13	0.20
$F1$-NCW	0.54	0.92	0.91	0.91	**0.93**	0.91

Table 4: Summary of performance metrics (m:mean and s:standard deviatio) across datasets. The additional experiments were conducted to investigate whether the traditional models that were trained using the features of the proposed method could produce stable results.

Dataset	$F1$-μ		$F1$-CW		$F1$-NCW		$F1$-NFS		$F1$-UFS		$F1$-CFS	
	m	s	m	s	m	s	m	s	m	s	m	s
CB	0.71	0.00					0.83	0.00	0.00	0.00	0.02	0.03
CT20	0.64	0.02					0.83	0.00	0.00	0.00	0.02	0.00
CT21	0.66	0.04	0.35	0.01	0.88	0.00						
CT22	0.79	0.01	0.10	0.00	0.94	0.00						
ENV	0.76	0.00	0.10	0.00	0.93	0.00						
NLP4IF	0.68	0.02	0.21	0.00	0.89	0.00						

Additionally, we observed that our approach maintains stable performance even when applied to different models and datasets.

These findings highlight two primary advantages of the proposed method: improved cross-domain versatility and greater output stability.

Determining whether a claim is worth fact-checking is not a question that can be answered with a single word; rather, it involves a complex judgment that requires consideration of multiple factors. By decomposing this decision into a series of more interpretable subtasks using the structured checklist, and assessing check-worthiness based on their outcomes, the method facilitates more reliable judgments at each stage. This structured approach likely contributed to the overall effectiveness of the system.

Our method was effective in identifying not check-worthy claims, but its performance was limited when detecting check-worthy ones. This indicates that the criteria for determining check-worthiness may be far more complex than those based on checklist-style impressions. By refining the current responses and developing a more sophisticated framework, more accurate predictions may be achieved. Ultimately, this research aims to develop reliable computational tools for human fact-checking.

Acknowledgments

This work was partly supported by JST RISTEX #JPMJRS23L2, Tateishi Science and Technology Foundation Research Grant C #237018 and the Grants-in-Aid for Academic Promotion, Graduate School of Culture and Information Science, Doshisha University.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barrón-Cedeño, A.; Alam, F.; Chakraborty, T.; Elsayed, T.; Nakov, P.; Przybyła, P.; Struß, J. M.; Haouari, F.; Hasanain, M.; Ruggeri, F.; et al. 2024. The CLEF-2024 CheckThat! Lab: Check-Worthiness, Subjectivity, Persuasion, Roles, Authorities, and Adversarial Robustness. In *European Conference on Information Retrieval*, 449–458. Springer.

Das, A.; Liu, H.; Kovatchev, V.; and Lease, M. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2): 103219.

Elsayed, T.; Nakov, P.; Barrón-Cedeño, A.; Hasanain, M.; Suwaileh, R.; Da San Martino, G.; and Atanasova, P. 2019. Overview of the CLEF-2019 CheckThat!: Automatic Identification and Verification of Claims. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, LNCS. Lugano, Switzerland.

Gangi Reddy, R.; Chinthakindi, S. C.; Wang, Z.; Fung, Y.; Conger, K.; ELsayed, A.; Palmer, M.; Nakov, P.; Hovy, E.; Small, K.; and Ji, H. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6002–6018. Stroudsburg, PA, USA: Association for Computational Linguistics.

Gencheva, P.; Sofia University "St. Kliment Ohridski", Bulgaria; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; Koychev, I.; Qatar Computing Research Institute, HBKU, Qatar; Qatar Computing Research Institute, HBKU, Qatar; Qatar Computing Research Institute, HBKU, Qatar; and Sofia University "St. Kliment Ohridski", Bulgaria. 2017a. A context-aware approach for detecting worth-checking claims in political debates. In Mitkov, R.; and Angelova, G., eds., *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, 267–276. Varna, Bulgaria: Incoma Ltd. Shoumen, Bulgaria.

Gencheva, P.; Sofia University "St. Kliment Ohridski", Bulgaria; Nakov, P.; Màrquez, L.; Barrón-Cedeño, A.; Koychev, I.; Qatar Computing Research Institute, HBKU, Qatar; Qatar Computing Research Institute, HBKU, Qatar; Qatar Computing Research Institute, HBKU, Qatar; and Sofia University "St. Kliment Ohridski", Bulgaria. 2017b. A context-aware approach for detecting worth-checking claims in political debates. In Mitkov, R.; and Angelova, G., eds., *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, 267–276. Varna, Bulgaria: Incoma Ltd. Shoumen, Bulgaria.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A sur-

vey on automated fact-checking. *Trans. Assoc. Comput. Linguist.*, 10: 178–206.

Hassan, N.; Li, C.; and Tremayne, M. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1835–1838.

Jaradat, I.; Gencheva, P.; Barrón-Cedeño, A.; Màrquez, L.; and Nakov, P. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. *arXiv preprint arXiv:1804.07587*.

Jha, R.; Motwani, E.; Singhal, N.; and Kaushal, R. 2023. Towards automated check-worthy sentence detection using Gated Recurrent Unit. *Neural Comput. Appl.*, 35(15): 11337–11357.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Leng, Y. 2024. Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases? *Available at SSRN 4705130*.

Majer, L.; and Šnajder, J. 2024. Claim check-worthiness detection: How well do LLMs grasp annotation guidelines? In Schlichtkrull, M.; Chen, Y.; Whitehouse, C.; Deng, Z.; Akhtar, M.; Aly, R.; Guo, Z.; Christodoulopoulos, C.; Cocarascu, O.; Mittal, A.; Thorne, J.; and Vlachos, A., eds., *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 245–263. Stroudsburg, PA, USA: Association for Computational Linguistics.

Nakov, P.; Da San Martino, G.; Elsayed, T.; Barrón-Cedeño, A.; Míguez, R.; Shaar, S.; Alam, F.; Haouari, F.; Hasanain, M.; Babulkov, N.; Nikolov, A.; Shahi, G. K.; Struß, J. M.; and Mandl, T. 2021. The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. In *Proceedings of the 43rd European Conference on Information Retrieval*, ECIR '21, 639–649. Lucca, Italy.

Nenno, S. 2024. Is checkworthiness generalizable? Evaluating task and domain generalization of datasets for claim detection. *Neural computing & applications*.

Neumann, T.; Lee, S.; De-Arteaga, M.; Fazelpour, S.; and Lease, M. 2024. Diverse, but Divisive: LLMs Can Exaggerate Gender Differences in Opinion Related to Harms of Misinformation. *arXiv preprint arXiv:2401.16558*.

Neumann, T.; and Wolczynski, N. 2023. Does AI-assisted fact-checking disproportionately benefit majority groups online? In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 480–490. New York, NY, USA: ACM.

Patwari, A.; Goldwasser, D.; and Bagchi, S. 2017. Tathya: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2259–2262.

Quelle, D.; and Bovet, A. 2024. The perils and promises of fact-checking with large language models. *Frontiers in Artificial Intelligence*, 7: 1341697.

Rastogi, S.; and Bansal, D. 2023. A review on fake news detection 3T's: typology, time of detection, taxonomies. *International Journal of Information Security*, 22(1): 177–212.

Schmitt, V.; Csomor, B. P.; Meyer, J.; Villa-Areas, L.-F.; Jakob, C.; Polzehl, T.; and Möller, S. 2024. Evaluating Human-Centered AI Explanations: Introduction of an XAI Evaluation Framework for Fact-Checking. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, 91–100.

Sehat, C. M.; Li, R.; Nie, P.; Prabhakar, T.; and Zhang, A. X. 2024. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1): 1–36.

Shaar, S.; Alam, F.; Da San Martino, G.; Nikolov, A.; Zaghouani, W.; Nakov, P.; and Feldman, A. 2021. Findings of the NLP4IF-2021 Shared Task on Fighting the COVID-19 Infodemic and Censorship Detection. In *Proceedings of the Fourth Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, NLP4IF@NAACL' 21. Online: Association for Computational Linguistics.

Teramoto, Y.; Komamizu, T.; Matsushita, M.; and Hatano, K. 2024. Feature Extraction for Claim Check-Worthiness Prediction Tasks Using LLM. In *Information Integration and Web Intelligence - 26th International Conference, iiWAS 2024, Bratislava, Slovak Republic, December 2-4, 2024, Proceedings, Part I*, 53–58.

Wang, Y.; Cai, Y.; Chen, M.; Liang, Y.; and Hooi, B. 2023. Primacy Effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 108–115.

Paper Checklist to be included in your paper

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, our work focuses on improving the prioritization process in fact-checking systems and does not involve any personally identifiable information or sensitive user data.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, we explicitly describe our contributions in the abstract and introduction and ensure consistency throughout the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we justify our use of feature-based classification and LLM prompting for claim check-worthiness prediction.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we acknowledge that the distribution of claims in existing fact-checking datasets may reflect biases related to topic prevalence or annotator demographics.
   (e) Did you describe the limitations of your work? Yes, we acknowledge that the distribution of claims in existing fact-checking datasets may reflect biases related to topic prevalence or annotator demographics.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we describe limitations related to annotation subjectivity and domain transferability
   (g) Did you discuss any potential misuse of your work? Yes, we highlight that overly relying on automatic check-worthiness predictions without human oversight may introduce bias or overlook minority viewpoints.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we ensure reproducibility through public code release and document data usage details
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them?
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results?
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? No
   (e) Did you address potential biases or limitations in your theoretical framework? We did it in the Section of Related Work
   (f) Have you related your theoretical results to the existing literature in social science? We did it in the Section of Related Work
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? No
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? No
   (b) Did you include complete proofs of all theoretical results? No
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? We did it in Abstract and the Section of Dataset
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? No
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? No
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? No
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see **?**)? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes

(d) Did you discuss how data is stored, shared, and deidentified? Yes