Benchmarking Check-Worthiness Models on LLM Generated Claims

Charlie Roadhouse, Matthew Shardlow, Ashley Williams
Manchester Metropolitan University
Ormond Building, Lower Ormond Street, Manchester, M15 6BX
charlie.roadhouse@stu.mmu.ac.uk, M.Shardlow@mmu.ac.uk, Ashley.Williams@mmu.ac.uk

Abstract

The proliferation of large language models (LLMs) has significantly increased the potential for automated dissemination of disinformation, necessitating robust systems for check-worthiness detection. However, existing models are primarily trained on human claims, leaving their performance on machine-generated text largely unexplored. In this paper, we benchmark encoder models (BERT and RoBERTa) and industry accessible tools (ClaimBuster) against LLM-paraphrased claims across three stylistic categories: syntactic restructuring, syntactic complexity and lexical informality. Our results indicate a consistent performance degradation on synthetic claims, particularly on complex and informal claims. We demonstrate that adversarial training significantly improves model resilience, with RoBERTa achieving F1-score gains up to +5.22 on the CheckIt dataset. Finally, SHAP analysis reveals that while base models rely on narrow syntactic heuristics such as active voice, robust models learn to anchor their prediction on core factual entities. These findings highlight the necessity of stylistic-aware training to maintain fact-checking efficacy in an increasingly LLM-populated information landscape.

Keywords: Fact-Checking, Check-Worthiness Detection, Adversarial Training, Disinformation

1. Introduction

The rise of generative AI methods utilising large language models (LLMs) to produce text has become widespread. The increase in LLM agency has led to their misappropriation for malicious purposes, where models are prompted to generate harmful or misleading content (Pan et al., 2023). Malicious actors leverage this generated text as disinformation, spreading it across online platforms. This proliferation of disinformation has a direct negative real-world impact across nations and socio-political topics (Aïmeur et al., 2023).

This threat has spurred research into the detection and mitigation of disinformation (Su et al., 2020). While these efforts show promise, they are yet to show their efficacy in real-world situations. Currently, fact-checking remains the most effective strategy for mitigating disinformation (Graves, 2017). The fact-checking process begins by identifying and assessing whether a claim is check-worthy, followed by evidence gathering to provide a final justification (Das et al., 2023). Due to the sheer volume of claims online, journalists can no longer manually incorporate this fact-checking process into their workflows, necessitating the work of dedicated fact-checking organisations such as PolitiFact[1], FullFact[2] & FactCheck.org[3].

While effective, these organisations face significant scalability and latency issues. Consequently, research has pivoted toward automated check-worthiness detection (Guo et al., 2022). These systems leverage AI models to automatically prioritise claims for review, removing the bottleneck of manual selection and significantly scaling number of claims processed (Graves, 2018). Such systems are now deployed in production workflows such as FullFact's internal AI suite[4] or ClaimBuster which is accessible through their API (Hassan et al., 2017)

The problem with current methods comes from their reliance on human annotated training data, often exhibiting low generalisation across topics, which can lead to claims incorrectly being identified within emerging events (Nenno, 2024). This reliance makes them susceptible to LLM-generated claims, which may contain linguistic nuances and structural discrepancies that differ from human-written claims. To understand the impact of LLM-generated content on model robustness, we propose a systematic study comparing state-of-the-art check-worthiness strategies against both human-authored and LLM-paraphrased claims. We then incorporate the LLM-generated content into the training of a separate model to adversarially train a model. All experiments are carried out on English claims only. We address the following research questions (RQ):

- **RQ1** - To what extent does a check-worthiness classification models' performance change when assessing LLM generated claims?
- **RQ2** - Does adversarial training using LLM generated claims improve the models performance on human written and LLM written claims?

[1] https://www.politifact.com/
[2] https://fullfact.org/
[3] https://www.factcheck.org/
[4] https://fullfact.org/ai/

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 34–44
12 May 2026.

- **RQ3** - Do models learn different linguistic features between the human and LLM generated claims?

2. Background

The impact of fake news is well documented to have influenced major socio-political events (Henkel, 2021; Chen et al., 2021; Cazzamatta, 2025). Given that fake news spans diverse topics, countries and modalities, modern research necessitates a more fine-grained taxonomy. Wardle and Derakhshan (2017) categorise information disorder into three sub-categories based on intent and truthfulness: Misinformation, Disinformation and Malinformation. Our study is grounded in this framework, specifically addressing the threat of automated disinformation scaling. By evaluating how models respond to stylistic circumvention, we provide insights into the resilience of check-worthiness systems against intentionally machine generation disinformation.

This categorisation allows researchers to better identify the motivations of authors and the susceptibilities of target audience (Bragazzi and Garbarino, 2024). While these categories provide a clear theoretical framework, there is a significant difference, with related natural language processing (NLP) tasks such as rumour detection (Bondielli and Marcelloni, 2019), propaganda detection (Da San Martino et al., 2020), credibility assessment (Srba et al., 2025) and fact-checking (Guo et al., 2022).

In the NLP community, early misinformation detection focused on machine learning models leveraging hand-crafted linguistic features for classification, as outlined in the survey by Su et al. (2020), which cover systems and datasets from 2000 - 2018. However, the introduction of transformer-based architectures, particularly BERT (Devlin et al., 2019), shifted the state-of-the-art toward encoder-based strategies that benefit from task-specific fine-tuning. Beyond textual content, recent research has also explored network topology and propagation patterns to identify misleading content (Zhou and Zafarani, 2019).

Despite these advancements, manual fact-checking remains the gold-standard for mitigation. To address the scalability issues of human fact-checkers, automated check-worthiness systems are employed to triage claims. In NLP, this is primarily framed as either a classification task (assigning binary or ordinal label) or a ranking task (prioritising claims based on perceived check-worthiness) (Guo et al., 2022).

Current state-of-the-art check-worthiness models leverage large-scale transformer encoders. For example, systems like ClaimBuster (Hassan et al., 2017) evolved from Support Vector Machines (SVM) to BERT-based architectures. These models often incorporate auxiliary information, such as named entity recognition (NER) and user metadata, to bolster performance (Ahmed et al., 2021).

However, encoder-based models often struggle with domain-shifting and emerging topics not present in the training data (Nenno, 2024). While LLMs have demonstrated impressive zero-shot capabilities in check-worthiness detection, issues such as hallucination and inconsistent reasoning mean that fine-tuned encoders remain the preference (Majer and Šnajder, 2024). Wright and Augenstein (2020) demonstrated that LLM-based paraphrasing is highly effective for increasing training data volume. Yet, a significant gap remains; while these studies use LLMs to support model training, they do not sufficiently investigate model performance in adversarial settings where LLMs are used to generate deceptive variations. This research aims to bridge the gap between LLM-augmented training and model robustness against synthetic stylistic shifts.

3. Methodology

3.1. Dataset Selection and Pre-Processing

To investigate the robustness of check-worthiness models against both human-authored and LLM-generated claims, two check-worthiness datasets are selected:

- **CheckIt (Sundriyal et al., 2026)**: A fine-grained assessment of claims from twitter. While annotating for a final check-worthiness label, data was also annotated on fine-grained labels. These comprised of if the claim is verifiable, publicly interesting, potentially harmful, misleading, of interest to the government and whether a fact-checker should check the claim.

- **CheckThatLab! 2024 (Hasanain et al., 2024)**: Made up event specific topics including COVID-19, vaccines, Gazza/Palestine conflict and climate change. Spanning both tweets, debates and speeches, framed as a binary classification task.

All code, models and data are released for reproducibility purposes[5].

3.2. Experimental Setup

We train two tracks of model: a base model (human only claims) and a robust model using adversarial training. We frame this as adversarial because the

[5]https://github.com/chroadhouse/Benchmarking-Check-Worthiness-Models-on-LLM-Generated-Claims

LLM-paraphrased variants act as stylistic perturbations designed to break the models resilience on surface level heuristics while keeping the semantic ground truth.

For our base and robust models, we select two transformer-based encoders. We also use a zero-shot LLM and publicly deployed state-of-the-art check-worthiness system for benchmarking. The models are outlined:

- **BERT** (Devlin et al., 2019): A standard baseline encoder widely utilised in prior check-worthiness shared tasks (Alam et al., 2023).
- **RoBERTa** (Liu et al., 2019): A robustly optimised variant of BERT. Its pre-training objective may offer enhanced resilience against stylistic variations in synthetic text.
- **Llama 3.1:8B**[6]: Employed as a zero-shot baseline to assess the check-worthiness detection capabilities of current general-purpose LLMs. Our prompt for classification is shown in Figure 1.
- **ClaimBuster** (Hassan et al., 2017): A leading industry-standard system providing a baseline for live fact-checking API[7] performance, leveraging a BERT-based architecture.

System Prompt
ROLE: You are a professional Fact-Checking Editor.
TASK: Analyse the following text and decide if it is **CHECK-WORTHY**.

CRITERIA FOR YES:
1. It contains a **verifiable factual claim** (dates, numbers, events, causal relationships).
2. It is of **public interest or potential harm** (not just a personal opinion or vague complaint).
3. It is **NOT standard health advice** (e.g., 'Wash your hands' is NO).

Input Format: `Text: {claim}`

Output Instruction: You must start your response with the decision tag.
Format: `<decision>YES</decision>` or `<decision>NO</decision>`

Figure 1: System prompt for zero-shot check-worthiness classification.

For the training, both encoder base models are trained on the CTL 24 training data, selected for cross domain coverage across social media, speeches and debates (Hasanain et al., 2024). The encoder models were trained for 10 epochs or until convergence, with an early stopping patience of 3. The models were optimised using AdamW (Loshchilov and Hutter, 2017) with a weight decay of 0.01 and a learning rate of 2×10^{-5}. We use the same parameters but also include LLM-paraphrased variants of the claims for the adversarial training. When classifying, we use a greedy decoding strategy for the Llama 3.1:8B model.

[6] https://huggingface.co/meta-llama/Llama-3.1-8B

[7] http://idir.uta.edu/claimbuster

Base Paraphrase Prompt (Syntactic Restructuring)

You are a precise fact-checking rewriter. Paraphrase the text below by changing the sentence structure (e.g., active to passive voice) while keeping all names, dates, and technical terms EXACTLY as written. Do not add any commentary.
Input: `Text: {text} <output>`

Complex Style Prompt (Syntactic Density)

You are a sophisticated academic editor. Rewrite the text below into a high-complexity, formal sentence using subordinate clauses and nominalization. You must retain all names, dates, and technical terms EXACTLY as written. Do not add any commentary.
Input: `Text: {text} <output>`

Informal Style Prompt (Lexical Informality)

You are a social media user. Rewrite the text below to sound like a natural, conversational post using contractions and informal phrasing. You must keep all specific names, dates, and statistics EXACTLY as written. Do not use hashtags or emojis. Do not add any commentary.
Input: `Text: {text} <output>`

Figure 2: Design of the LLM prompt templates used for generating the three stylistic paraphrases (Syntactic Restructuring, Syntactic Density, and Lexical Informality).

3.3. Generation of LLM Content

We leverage Llama 3.1:8B-Instruct to generate paraphrased candidates for the CheckThatLab 2024 and CheckIt datasets, using an auto regressive strategy with nucleus sampling (Top-P) set to 0.9, with a temperature of 0.8 and a top-k of 40. To ensure linguistic variety preventing structural loops, a repetition of 1.1 is applied.

We experiment with three zero-shot prompting strategies (see Figure 2): Standard, Informal and

Complex. For each variant, the model is strictly instructed to maintain all named entities to minimise factual hallucination. Table 1 illustrates an original human-authored claim alongside its three corresponding synthetic variants, showing the linguistic differences while ensuring entities are not replaced.

Ensuring label preservation is critical when generating synthetic data; a paraphrase that alters the factual content of the claim would invalidate the original check-worthiness labels. Consequently, we assess the linguistic and semantic drift between the human reference and LLM-generated candidates using a suite of metrics. Given that factual integrity is more vital in this domain than absolute semantic overlap (Wright and Augenstein, 2020), our validation framework prioritises entity preservation and semantic stability.

To extract the necessary features for these metrics we utilise a RoBERTa model fine-tuned for NER on the TweetNER7 dataset (Antypas et al., 2023) and a second RoBERTa model fine-tuned on the Tweet-Eval benchmark for sentiment analysis (Barbieri et al., 2020). The performance of the generation process is then quantified using the following metrics, with results summarised in Table 2.

- **BLEURT** (Sellam et al., 2020): An evaluation metric that leverages an encoder model trained on perturbed sentence pairs and then trained on multiple semantic preservation metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) & BERTScore (Zhang et al., 2019), all of which are popular N-gram overlap methods.
- **Mutual Implication Score (MIS)** (Babakov et al., 2022): This method is developed specifically to understand the similarity between a reference piece of text and a paraphrased candidate. It leverages a natural language inference model, treating the paraphrasing as bidirectional entailment.
- **Cosine Similarity of Entity Distributions**: We measure the preservation of factual subjects by comparing the frequency of entity groups (e.g., persons, organisations, locations) between original and paraphrased claims. We identify named entities and represent each claim by a vector (category count). By calculating the cosine similarity (Büyüktopaç and Acarman, 2019) between these vectors, we quantify how well the LLM maintains the original subjects, ensuring stylistic shifts do not alter the factual composition of the claim.
- **Jaccard Similarity** (Niwattanakul et al., 2013): The Jaccard similarity is used to measure the difference in the extracted named entities. This is crucial to ensure that the same names, dates and locations are mentioned, even with a change of phrasing.
- **Sentiment** Δ: The sentiment score for both the reference and candidate claim are extracted and then the difference is calculated to see the sentiment drift between the two claims.
- **Length** Δ: The difference in characters between the original reference and candidate claim.

The results in Table 2 demonstrate high entity group consistency between paraphrased and human claims. While the lower BLEURT and MIS scores suggest significant stylistic and structural divergence, the stability of the entity metrics confirms that the factual subjects remain intact during the transformation.

To ensure that the LLM did not hallucinate in the generation of paraphrased claims, we randomly select 50 samples from each of the three prompts from the CTL 2024 dataset and assess semantic similarity between the human and LLM-generated pair. The primary investigator annotated for semantic preservation using a binary label and then used a predefined confidence definition by Mu et al. (2023) to rate the confidence of the decision between 1-5. The results of the preservation annotation are presented in Table 3. They show a high semantic preservation with a mean confidence over 4.

4. Results

Table 4 presents the performance of the fine-tuned transformer models, ClaimBuster and zero-shot LLM. Performance is evaluated across three benchmark datasets using weighted Precision (P), Recall (R) and F1-score. The results compare the efficacy on original human-authored claims against their LLM-paraphrased counterparts.

The experimental results indicate that fine-tuned transformer models outperform ClaimBuster and zero-shot LLM across both the CheckIt and CTL 2024 data splits. All four models experienced a performance degradation when evaluated on the LLM paraphrased claims, specifically in the Complex and Informal settings. Notably, while the ClaimBuster model remained competitive on the CTL 2024 test split, the model struggled on the CheckIt dataset.

4.1. Robust Model Performance

The results of the adversarial training are detailed in Table 5. Where the difference between the original and robust encoder models are calculated for each dataset split and claim type. Table 5 highlights a significant performance boost for the robust

Claim Type	Claim Text	Linguistic Changes
Human (Original)	"We've run up more debt in the last eight years than under all the presidents from George Washington to Jimmy Carter combined."	*Baseline:* Natural, conversational political speech.
Simple Paraphrase	"More debt has been run up in the last eight years than under all the presidents from George Washington to Jimmy Carter combined."	**Syntactic Shift:** Conversion from active to passive voice; maintains original vocabulary.
Informal Style	"We've racked up more debt in the last eight years than all the presidents from George Washington to Jimmy Carter combined."	**Lexical Shift:** Substitution of "run up" with the colloquial phrasal verb "racked up."
Complex Style	"The accumulation of debt over the preceding eight years has exceeded, in magnitude, the total indebtedness incurred during the collective administrations of all presidents from George Washington to Jimmy Carter."	**Semantic Over-Formalisation:** Massive expansion of word count through bureaucratic synonyms (e.g., "accumulation," "indebtedness," "collective administrations").

Table 1: Comparison of Human Claim and LLM-Generated paraphrases across different styles.

			Semantic Analysis		Entity Integrity		Style	Length
Dataset	**Split**	**Category**	**BLEURT**	**MIS**	**Jaccard**	**Cosine**	**Sent. Δ**	**Len Δ**
CT24	Validation	Base	0.3734	0.2543	0.9383	0.9608	-0.0072	9.76
		Informal	0.3732	0.2555	0.9547	0.9738	0.0023	8.17
		Complex	0.3619	0.2774	0.9210	0.9599	-0.0108	90.45
	Testing	Base	0.3698	0.2535	0.9283	0.9633	-0.0072	9.62
		Informal	0.3706	0.2533	0.9605	0.9789	-0.0052	6.94
		Complex	0.3580	0.2763	0.9238	0.9538	-0.0290	96.95
CheckIt	Validation	Base	0.3526	0.2517	0.7493	0.8447	-0.0109	12.24
		Informal	0.3517	0.2533	0.7433	0.8387	-0.0056	9.38
		Complex	0.3355	0.2914	0.7110	0.8167	-0.0363	132.52
	Testing	Base	0.3530	0.2524	0.7622	0.8547	-0.0137	11.87
		Informal	0.3525	0.2534	0.7628	0.8629	-0.0006	9.13
		Complex	0.3357	0.2932	0.7309	0.8364	-0.0190	128.97

Table 2: Comprehensive evaluation of claim paraphrasing categorised by Dataset, Split, and Stylistic Category. Results indicate consistency in meaning preservation across validation and testing sets.

Prompt Style	Samples (n)	Semantic Preservation (%)	Mean Confidence (1–5)
Informal	50	88.0%	4.14
Complex	50	92.0%	4.22
Base (Rephrased)	50	84.0%	4.08
Total	**150**	**88.0%**	**4.15**

Table 3: Manual validation of LLM-generated paraphrases for semantic preservation and annotator confidence.

encoder models, particularly when handling LLM-paraphrased claims. The most notable improvements occur within the CheckIt dataset, where RoBERTa demonstrates an F1-score increase of + 5.22 on the validation split and + 3.61 on the testing split. BERT similarly shows consistent gains across CheckIt informal and complex claim types. Conversely, the CTL 2024 results present a more varied outcome; while validation scores remain stable or improve, some testing splits experience marginal declines.

4.2. SHAP Analysis

To investigate the linguistic shifts influencing model decisions, we utilised SHAP analysis (Mosca et al., 2022) to identify high attribution tokens for classification. We select the base and robust variants of RoBERTa to investigate classification, due to being the most consistent model.

As shown in Table 6, we measure the intersection

Dataset	Claim Type	Model	Validation Split			Testing Split		
			P	R	F1	P	R	F1
CheckIt	Human	BERT	81.36	81.53	80.99	78.56	78.94	78.46
		RoBERTa	80.86	81.08	80.55	80.47	80.63	79.98
		Llama 3.1	53.41	63.06	57.67	59.40	63.74	58.00
		ClaimBuster	67.43	58.67	59.08	67.76	57.43	57.49
	Base	BERT	98.34	77.93	78.09	76.79	76.46	76.60
		RoBERTa	77.62	75.90	76.32	77.33	76.35	76.65
		Llama 3.1	63.11	65.88	59.37	60.77	64.64	58.22
		ClaimBuster	67.69	63.06	63.97	66.00	61.37	62.20
	Informal	BERT	79.03	76.24	76.73	77.16	74.44	74.96
		RoBERTa	79.66	77.48	77.91	77.83	76.58	76.92
		Llama 3.1	62.50	65.54	60.69	67.37	57.77	58.02
		ClaimBuster	67.37	57.77	58.02	65.14	54.84	54.80
	Complex	BERT	73.07	73.54	73.22	72.74	73.20	72.89
		RoBERTa	74.58	73.09	73.51	72.56	71.28	71.69
		Llama 3.1	58.75	64.19	55.13	64.59	64.64	64.61
		ClaimBuster	64.59	64.64	64.61	65.30	64.53	64.85
CTL 2024	Human	BERT	97.86	97.87	97.86	87.82	87.98	86.56
		RoBERTa	97.29	97.29	97.25	87.34	87.10	86.05
		Llama 3.1	88.19	77.42	79.18	80.33	68.33	70.29
		ClaimBuster	98.68	98.64	98.65	82.84	83.68	82.84
	Base	BERT	97.01	96.90	96.93	86.47	86.80	86.56
		RoBERTa	97.31	97.29	97.25	87.53	87.68	86.99
		Llama 3.1	88.26	76.65	78.48	80.93	66.28	68.27
		ClaimBuster	97.24	97.00	97.05	83.95	84.16	84.04
	Informal	BERT	97.18	97.19	97.18	87.46	87.68	87.06
		RoBERTa	97.47	97.48	97.49	87.19	87.10	86.13
		Llama 3.1	88.92	80.43	81.90	81.07	71.55	73.32
		ClaimBuster	97.73	97.67	97.67	32.53	83.28	82.58
	Complex	BERT	95.13	94.96	95.02	86.80	87.10	86.89
		RoBERTa	96.88	96.90	96.87	87.19	87.10	86.13
		Llama 3.1	87.69	74.90	76.88	78.76	64.81	66.92
		ClaimBuster	94.98	94.19	94.36	85.68	85.63	85.66

Table 4: Comparative performance of check-worthiness models grouped by claim type.

between the tokens by SHAP attribution and named entities within each claim. For the human-authored claims in CheckIt test set, the robust RoBERTa model demonstrates a significantly higher alignment with factual entities compared to the base model. Suggesting that adversarial training encourages the model to anchor the check-worthiness decision to these verifiable subjects such as names, dates and organisations.

5. Discussion

5.1. RQ1: Sensitivity to Synthetic Phrasing

The experimental results in Table 4 confirm that state-of-the-art check-worthiness models are sensitive to stylistic variations introduced by LLM-based paraphrasing. BERT and RoBERTa exhibit a notable performance degradation when processing synthetic claims within the CheckIt dataset, being most pronounced in the complex and informal categories. However, performance remains relatively stable on the CTL 2024 test sets. This

Dataset	Model	Claim Type	Validation Split (△)			Testing Split (△)		
			P	R	F1	P	R	F1
CheckIt	BERT	Human	+0.27	+0.34	+0.66	+1.77	+1.69	+1.88
		Base	-17.88	+2.70	+2.44	+2.87	+3.38	+3.13
		Informal	-0.37	+1.80	+1.52	+0.70	+2.25	+2.06
		Complex	+2.21	+1.12	+1.67	+2.93	+1.57	+2.18
	RoBERTa	Human	+1.76	+1.69	+1.79	-1.04	-0.90	-0.84
		Base	+4.38	+6.19	+5.22	+3.15	+4.39	+3.61
		Informal	+2.00	+4.16	+3.74	+0.05	+1.46	+1.03
		Complex	+2.34	+4.27	+3.04	+1.54	+3.49	+2.23
CTL 2024	BERT	Human	+0.10	+0.10	+0.10	-0.75	-1.18	-0.88
		Base	+0.39	+0.48	+0.46	-2.01	-2.05	-3.05
		Informal	0.00	0.00	-0.02	-2.58	-2.64	-3.28
		Complex	+2.14	+2.33	+2.25	-0.03	-0.30	-1.03
	RoBERTa	Human	+0.13	+0.09	+0.15	0.00	+0.58	+1.27
		Base	-0.97	-1.26	-1.15	+0.70	+0.30	+1.09
		Informal	-0.90	-1.07	-1.03	+0.79	+1.17	+1.90
		Complex	-1.93	-2.62	-2.44	-0.47	-1.18	+0.07

Table 5: Performance delta (Δ = Robust − Original) for BERT and RoBERTa models across datasets.

Dataset	Model	Human (%)	Base (%)	Complex (%)	Informal (%)
CheckIt Val	Base	15.14	12.76	9.56	14.67
	Robust	13.47	11.77	7.90	15.38
CheckIt Test	Base	15.69	9.72	6.93	13.43
	Robust	**20.45**	**17.28**	**9.65**	**15.53**
CT24 Val	Base	6.76	6.88	7.41	8.18
	Robust	8.29	7.61	6.82	11.20
CT24 Test	Base	5.11	6.75	4.90	4.30
	Robust	7.55	7.73	5.75	7.52

Table 6: Percentage of overlap between SHAP attribution tokens and named entities. This metric indicates the extent to which the model anchors its check-worthiness prediction on core factual subjects vs. stylistic noise.

suggest that while models are sensitive to stylistic shifts, the impact is domain-dependent, with event-specific datasets potentially offering higher inherent resilience.

Crucially, the semantic validation metrics in Table 2 suggest that this decline is not caused by a loss of factual or semantic information. The MIS and cosine similarity remain high across all scores, indicating that the core entities and claims remain intact. Instead the models appear to rely on specific structural patterns inherent to human social media discourse. When the LLM restructures these claims, it effectively masks the check-worthiness signals that the encoder models were fine-tuned to detect.

The relative stability of the zero-shot LLM results suggest a state of distributional parity between the model evaluating the claims and the one generating the paraphrased variants. Unlike the encoders, the LLM exhibits a lack of dependency on surface level heuristics, which are acquired in the fine-tuning on human-authored corpora. While this leads to a more consistent performance across claim types, it also highlights domain insensitivity; the LLM appears unable to prioritise the contextual urgency and pragmatic cues that signify check-worthiness in the organic human discourse.

5.2. RQ2: Risks of Adversarial Training

The performance deltas presented in Table 5 indicate that while adversarial training is a viable strategy for improving robustness, it carries significant risks related to distributional interference and architectural sensitivity.

Given that the encoder models were trained ex-

clusively on the CTL 2024 training data, the results on the CheckIt datasets serve as a measure of stylistic transfer. Interestingly, for the CheckIt dataset, both BERT and RoBERTa show consistent improvements across almost all claim types. RoBERTa, in particular, achieves an F1-score increase of + 5.22 (Validation) and + 3.61 (Testing) on base paraphrases. This suggests that exposing the model to the synthetic data during its training on the CTL 2024 domain provides a generalisable stylistic edge that transfers effectively to the CheckIt data. However, the primary risk of this approach is evidenced in the CTL 2024 testing split, the same domain used for training. Here, we observe a distinct difference between architectures. While we observe a marginal gain in validation, it suffers a significant performance collapse in the test set. Specifically, the robust BERT variant experimented an F1-score decline of -3.05 on base and -3.28 on informal claims. Because the model was trained on the human and synthetic variants of the CTL 2024 data, these negative delta suggest that for BERT, the synthetic samples acted as a distraction. The model likely began to over fit to the paraphrased claims, losing its ability to generalise to the subtle pragmatic marks of check-worthiness.

In contrast, RoBERTa demonstrates an improvement on the CTL 2024 test set, with an increase of +1.27 on human claims and up to +1.9 on synthetic variants. This suggests that RoBERTa's robust optimisation and pre-training objective allow it to reconcile the divergent stylistic features of human social media discourse and synthetic LLM generated text more effectively than BERT.

While adversarial training significantly bolsters resilience on the CheckIt dataset, the marginal performance drop in the CTL 2024 test set suggests a degree of noise from the adversarial training. To ensure practical reliability in systems, future work should investigate curriculum learning (Wang et al., 2022), where the model is stabilised on the human-authored claims before then being exposed to the difficult synthetic claims.

5.3. RQ3: Feature Shift and Attribution

The systematic misclassification identified through qualitative error logs suggest that the fine-tuned encoders rely on narrow syntactic heuristics rather than semantic representations of check-worthiness. A primary failure is the models dependence on active voice. Our qualitative analysis shows that when LLMs shift a human claim to the passive voice, the models frequently flip from a correct to incorrect classification. Indicating that the models have internalised a claim-like syntax for verifiable intent, failing to recognise the same factual signal when restructured.

The degradation in the complex category (see Table 4) corresponds to a failure to process normalisation. When verbs are transformed into abstract noun phrases, the models often mis-classify the statement as not check-worthy. Suggesting that the encoder models conflate syntactic simplicity and prose with high priority factual assertions, while perceiving complex machine generated text as editorial or descriptive noise.

The Informal results reveal a persistent reliability heuristic. The addition of colloquialisms often trigger a negative prediction for claims previously identified as check-worthy in their human form. This implies that fine-tuning on organic human corpora leads models to associate standard formal registers with high priority verifiable claims and informal registers with subjective, low-priority claims. While the robust RoBERTa model begins to reconcile these features, the persistence of these errors suggest that even adversarial training struggles to decouple the need to check stylistic wrappers.

5.4. Future Work

While adversarial training improves resilience, several avenues remain for enhancing check-worthiness systems.

Categorical Information Disorder: Current models treat check-worthiness as a monolithic label. Future work should develop multi-label datasets incorporating the Wardle and Derakhshan (2017) framework. Distinguishing between misinformation (unintentional) and disinformation (malicious) would allow for more nuanced triage priorities, where a model prioritises claims not just by check-worthiness, but by potential harm and authorial intent.

Check-Worthiness Specific Adversarial Attacks: As practitioners increasingly deploy automated suites, malicious actors may weaponise LLMs to generate low-signal disinformation. Future research should investigate a broader spectrum of attacks, including homoglyphs (Roadhouse et al., 2024), character-level perturbations (Morris et al., 2020), and word shuffling (Lewoniewski et al., 2024). Developing robust training loops that simulate these specific evasion tactics is essential for maintaining defence in an adversarial information landscape.

Fully Synthetic Dataset Generation: Our paraphrasing approach demonstrates that models can identify LLM-restructured claims. A natural next step is the generation of fully synthetic datasets to alleviate the bottleneck of human annotation. However, this requires rigorous experimentation to ensure that synthetic claims reflect the pragmatic

nuances of human discourse without introducing systematic model hallucinations or reinforcement of existing training biases.

6. Conclusion

In this study, we investigated the robustness of automated check-worthiness models against the emerging threat of LLM generated claims. Our findings demonstrate that state-of-the-art transformer encoders are highly sensitive to stylistic and syntactic shifts introduced by LLM paraphrasing. Qualitative and SHAP-based analysis confirm that these models rely on surface-level heuristics, such as active voice and standard formal registers, leading to misclassification when factual claims are restructured into passive or formalised prose.

We show that while adversarial training is a viable strategy for mitigating this sensitivity, its efficacy is architecture dependent. While RoBERTa demonstrated significant improvements across both human and synthetic claims, BERT exhibited signs of distributional interference, suggesting that more robust pre-training objectives are required to reconcile the stylistic gaps between human and LLM discourse. Ultimately, our work underscores the need for automated check-worthiness and fact-checking as a whole to remain a viable defence against disinformation, models must move beyond syntactic patterns and develop to a more grounded representation of factual priority.

Limitations

Despite the insights gained from this benchmarking study, several limitations must be acknowledged. First, our evaluation relies on LLM-paraphrased claims derived from human-authored ground truths rather than fully fabricated synthetic disinformation. While this ensures label preservation for experimental control, it may not capture the full spectrum of linguistic hallucination or logical inconsistencies present in completely machine generated narratives.

Furthermore, the synthetic dataset was generated using a single model. While this provided a controlled environment for testing specific stylistic shifts, different LLM architectures exhibit varying generation entropies and stylistic biases which may limit the generalisability of these findings. Additionally, our manual verification process, which served as a primary hallucination check to ensure the LLM maintained factual integrity during paraphrasing, which was conducted by a single investigator. While the results showed high annotator confidence, the absence of multiple annotators precludes the calculation of inter-annotator agreement.

Finally, our analysis is restricted to English claims within the socio-political and public health domains. The stylistic heuristics identified (e.g. active vs passive voice) may manifest differently in other languages or specialised domains such as legal or scientific fact-checking.

Ethical Considerations

The primary ethical concern regarding this research is the dual-use risk inherent in probing model vulnerabilities. By identifying the specific stylistic wrappers that allow claims to bypass automated detection, there is a potential for a malicious actor to weaponise these findings to generate adversarial disinformation. However, by documenting these weakness, they should be viewed as a prerequisite for developing robust adversarial resistant models.

7. Bibliographical References

Sajjad Ahmed, Klestia Balla, Knut Hinkelmann, and Flavio Corradini. 2021. Fact Checking: Detection of Check Worthy Statements Through Support Vector Machine and Feed Forward Neural Network. In *Advances in Information and Communication*, pages 520–535, Cham. Springer International Publishing.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Gautam Kishore Shahi, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Wajdi Zaghouani, et al. 2023. Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content. In *CLEF (Working Notes)*, pages 219–235.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. A large-scale computational study of content preservation measures for text style transfer and paraphrase generation. In *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics: Student Research Workshop, pages 300–321, Dublin, Ireland. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55.

Nicola Luigi Bragazzi and Sergio Garbarino. 2024. Understanding and combating misinformation: An evolutionary perspective. *JMIR infodemiology*, 4(1):e65521.

Onur Büyüktopaç and Tankut Acarman. 2019. Evaluation of cosine similarity feature for named entity recognition on tweets. In *International Conference on Man–Machine Interactions*, pages 125–135. Springer.

Regina Cazzamatta. 2025. Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries. *New Media & Society*, 27(11):6334–6358.

Emily Chen, Herbert Chang, Ashwin Rao, Kristina Lerman, Geoffrey Cowan, and Emilio Ferrara. 2021. Covid-19 misinformation and the 2020 us presidential election. *Harvard kennedy school misinformation review*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Yu Seunghak, Roberto Di Pietro, Preslav Nakov, et al. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization.

Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. 2023. The state of human-centered NLP technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

D. Graves. 2018. Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*.

L. Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture and Critique*, 10(3).

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics*, 10:178–206.

Maram Hasanain, Reem Suwaileh, Sanne Weering, Chengkai Li, Tommaso Caselli, Wajdi Zaghouani, Alberto Barrón-Cedeño, Preslav Nakov, and Firoj Alam. 2024. Overview of the clef-2024 checkthat! lab task 1 on check-worthiness estimation of multigenre content. In *25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024*, pages 276–286. CEUR Workshop Proceedings.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Imke Henkel. 2021. Ideology and disinformation: how false news stories contributed to brexit. *Politics of disinformation: The influence of fake news on the public sphere*, pages 79–90.

Włodzimierz Lewoniewski, Piotr Stolarski, Milena Stróżyna, Elżbieta Lewańska, Aleksandra Wojewoda, Ewelina Księżniak, and Marcin Sawiński. 2024. Openfact at checkthat! 2024: Combining multiple attack methods for effective adversarial text generation. In *CEUR Workshop Proceedings*, volume 3740.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Laura Majer and Jan Šnajder. 2024. Claim Check-Worthiness Detection: How Well do LLMs Grasp Annotation Guidelines? In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 245–263, Miami, Florida, USA. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations*, pages 119–126.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pages 4593–4603.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1052–1062.

Sami Nenno. 2024. Is checkworthiness generalizable? Evaluating task and domain generalization of datasets for claim detection. *Neural Computing and Applications*, 36(24):15165–15176.

Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of jaccard coefficient for keywords similarity. In *Proceedings of the international multiconference of engineers and computer scientists*, volume 1, pages 380–384.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Charlie Roadhouse, Matthew Shardlow, and Ashley Williams. 2024. Mmu nlp at checkthat! 2024: Homoglyphs are adversarial attacks. In *CLEF (Working Notes)*, pages 580–589.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7881–7892.

Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, et al. 2025. A survey on automatic credibility assessment using textual credibility signals in the era of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. 2020. Motivations, methods and metrics of misinformation detection: an nlp perspective. *Natural Language Processing Research*, 1(1):1–13.

Megha Sundriyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2026. Leveraging rationality labels for explainable claim check-worthiness. *IEEE Transactions on Artificial Intelligence*, 7(1):239–249.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A Survey on Curriculum Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Dustin Wright and Isabelle Augenstein. 2020. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 476–488, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter*, 21(2):48–60.