Generating Zero-shot Abstractive Explanations for Rumour Verification

Anonymous ACL submission

Abstract

The task of rumour verification in social media concerns assessing the veracity of a claim on the basis of conversation threads that result from it. While previous work has focused on predicting a veracity label, here we reformulate the task to generate model-centric free-text explanations of a rumour's veracity. The approach is model agnostic in that it generalises to any model. Here we propose a novel GNN-based rumour verification model. We follow a zero-shot approach by first applying post-hoc explainability methods to score the most important posts within a thread and then we use these posts to generate informative explanations using opinion-guided summarisation. To evaluate the informativeness of the explanatory summaries, we exploit the few-shot learning capabilities of a large language model (LLM). Our experiments show that LLMs can have similar agreement to humans in evaluating summaries. Importantly, we show explanatory abstractive summaries are more informative and better reflect the predicted rumour veracity than just using the highest ranking posts in the thread.[1]

1 Introduction

Evaluating misinformation on social media is a challenging task that requires many steps (Zubiaga et al., 2016): detection of rumourous claims, identification of stance towards a rumour, and finally assessing rumour veracity. In particular, misinformation may not be immediately verifiable using reliable sources of information such as news articles since they might not have been available at the time a rumour has emerged. For the past eight years, researchers have focused on the task of automating the process of rumour verification in terms of assigning a label of *true*, *false*, or *unverified* (Zubiaga et al., 2016; Derczynski et al., 2017). However, recent work has shown that while fact-checkers agree with the urgent need for computational tools for content verification, the output of the latter can only be trusted if it is accompanied by explanations (Procter et al., 2023).

Thus, in this paper, we move away from black-box classifiers of rumour veracity to generating explanations written in natural language (free-text) for why, given some evidence, a statement can be assigned a particular veracity status. This has real-world applicability particularly in rapidly evolving situations such as natural disasters or terror attacks (Procter et al., 2013), where the explanation for an automated veracity decision is crucial (Lipton, 2018). To this effect, we use a popular benchmark, the PHEME (Zubiaga et al., 2016) dataset, to train a rumour verifier and employ the conversation threads that form its input to generate model-centric explanation summaries of the model's assessments.

Atanasova et al. (2020), Kotonya and Toni (2020) and Stammbach and Ash (2020) were the first to introduce explanation summaries for fact-checking across different datasets. Kotonya and Toni (2020) provided a framework for creating abstractive summaries that justify the true veracity of the claim in the PUBHealth dataset, similarly to Stammbach and Ash (2020) who augment the FEVER (Thorne et al., 2018) dataset with a corpus of explanations. Atanasova et al. (2020) proposed a jointly trained system that produces veracity predictions as well as explanations in the form of extracted evidence from ruling comments on the LIAR-PLUS dataset (Alhindi et al., 2018). The approach in (Kotonya and Toni, 2020) results in explanatory summaries that are, however, not faithful to the model, while Atanasova et al. (2020) requires a supervised approach. Our goal is to create a novel zero-shot method for abstractive explanations that explain the rumour verification model's predictions. We make the following contributions:

- We introduce a zero-shot framework for generating abstractive explanations using opinion-

[1] A sample of generated explanations and code are provided. Colour-coded changes of the revised paper are in A. E.

guided summarisation for the task of rumour verification. To the best of our knowledge, this is the first time free-text explanations are introduced for this task.

- We investigate the benefits of using a gradient-based algorithm and a game theoretical algorithm to provide explainability.
- While our explanation generation method is generalisable to any verification model, we introduce a novel graph-based hierarchical approach.
- We evaluate the informativeness of several explanation baselines, including model-independent and model-dependent ones stemming from the highest scoring posts by providing them as input to a few-shot trained large language model. Our results show that our proposed abstractive model-centric explanations are more informative in 77% of the cases as opposed to 49% for all other baselines.
- We provide both human and LLM-based evaluation of the generated explanations, showing that LLMs achieve sufficient agreement with humans, thus allowing scaling of the evaluation of the explanatory summaries in absence of gold-truth explanations.

2 Related Work

Explainable Fact Checking Following the example of fact-checking organisations (e.g., Snopes, Full Fact, Politifact), which provide journalist-written justifications to determine the truthfulness of claims, recent datasets augmented with free-text explanations have been constructed: LIAR-PLUS (Alhindi et al., 2018), PubHealth (Kotonya and Toni, 2020), AVeriTeC (Schlichtkrull et al., 2023). A wide range of explainable outputs and methods have been proposed: theorem proofs (Krishna et al., 2022), knowledge graphs (Ahmadi et al., 2019), question-answer decompositions (Boissonnet et al., 2022; Chen et al., 2022), reasoning programs (Pan et al., 2023), deployable evidence-based tools (Zhang et al., 2021b) and summarisation (Atanasova et al., 2020; Kotonya et al., 2021; Stammbach and Ash, 2020; Kazemi et al., 2021; Jolly et al., 2022). We adopt summarisation as our generation strategy as it fluently aggregates evidence from multiple inputs and has been proven effective in similar works which we discuss next.

Explainability as Summarisation Atanasova et al. (2020) and Kotonya and Toni (2020) leveraged large-scale datasets annotated with gold justifications to generate supervised explanations for fact-checking, while Stammbach and Ash (2020) used few-shot learning on GPT-3 to create the e-FEVER dataset of explanations. Similar to (Stammbach and Ash, 2020), Kazemi et al. (2021) also leveraged a GPT-based model (GPT-2) to generate abstractive explanations, but found that that their extractive baseline, Biased TextRank, outperformed GPT-2 on the LIAR-PLUS dataset (Alhindi et al., 2018). Jolly et al. (2022) warn that the output of extractive explainers lacks fluency and sentential coherence, which motivated their work on unsupervised post-editing using the explanations produced by Atanasova et al. (2020). Our approach is different from the above as we derive our summaries from microblog content (as opposed to news articles as done by Atanasova et al. (2020); Stammbach and Ash (2020); Kazemi et al. (2021); Jolly et al. (2022), and only use the subset of posts relevant to the model's decision to inform the summary (rather than summarising the whole input as in (Kotonya and Toni, 2020; Kazemi et al., 2021). Moreover, we rely on a zero-shot generation approach without gold explanations, contrary to (Atanasova et al., 2020; Kotonya and Toni, 2020).

LLMs as evaluators Having generated explanatory summaries the question arises as to how to evaluate them at scale. LLMs have been employed as knowledge bases for fact-checking (Lee et al., 2020; Pan et al., 2023), as explanation generators for assessing a claim's veracity (Stammbach and Ash, 2020; Kazemi et al., 2021) and, as of recently, as evaluators in generation tasks. Most works focused on assessing the capability of LLM-based evaluation on summarisation tasks, either on long documents (Wu et al., 2023) or for low-resource languages (Hada et al., 2023). While there is work focusing on reducing positional bias (Wang et al., 2023b) and costs incurred (Wu et al., 2023) for using LLM-based evaluators, our evaluation is most similar to Liu et al. (2023); Shen et al. (2023); Chiang and Lee (2023), who study the extent of LLM-human agreement in evaluations of fine-grained dimensions, such as fluency or consistency. We believe we are the first to use an LLM-powered evaluation to assess the informativeness and faithfulness of explanations for verifying a claim.

3 Methodology

Our methodological approach (Figure 1) consists of three individual components: *i*) training a ru-

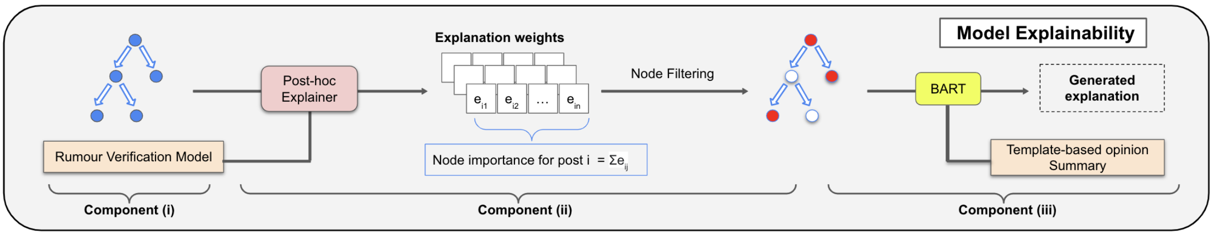

Figure 1: Framework of our proposed approach to obtain faithful generated explanations for the rumour verification model. It explains the process of explanation generation, where the weights from a model are passed through an explainer algorithm to identify important input nodes, which are then filtered and used in abstractive summarisation.

mour verification model; *ii*) using a post-hoc explainability algorithm; *iii*) generating summary-explanations. The approach to explanation generation is zero-shot and model-agnostic.

We demonstrate our approach on PHEME (Zubiaga et al., 2016), a widely used benchmark dataset for classifying social media rumours into either unverified, true or false. It contains conversation threads that cover 5 real-world events such as the Charlie Hebdo attack and the Germanwings plane crash. We adopt the same leave-one-out testing as previous works (Dougrez-Lewis et al., 2022) which favours real-world applicability as the model is tested on new events not included in the test data.

Task Formulation For a model trained on rumour verification $\mathcal{M}$, an attribution-based explanation method $\mathcal{E}$, and a rumourous conversation thread consisting of posts $\mathcal{T} = \{p_1, ...p_l\}$ with embeddings $\{x_1, ...x_l\} \subset \mathbb{R}^n$, we define the post importance as a function $f_{(\mathcal{M},\mathcal{E})} : \mathcal{T} \to \mathbb{R}$.

$$f_{(\mathcal{M},\mathcal{E})}(p_i) = \sum_{j=1}^{n} \mathcal{E}(\mathcal{M}, x_i)_j = \sum_{j=1}^{n} e_{ij} \quad (1)$$

where $e_i \in \mathbb{R}^n$ is the attribution vector for embedding x_i of post p_i such that each value e_{ij} corresponds to the weight of feature x_{ij} assigned by the explainer algorithm $\mathcal{E}$.

The summary is generated from the subset of posts that are most important for the model prediction, i.e., $\mathcal{I} = \{p_i \mid f_{(\mathcal{M},\mathcal{E})}(p_i) > 0\}$. Note a thread will contain posts that agree with the prediction (positive importance scores) and posts that disagree (negative importance scores).

3.1 Rumour Verification Model

Our explanation generation method is applicable to any rumour verification model, but here we chose an approach based on graph neural networks (See Figure 2), which caters for a flexible information structure combining information in the conversation thread with information about stance. This is the first time a GNN-based model enriched with stance has been proposed for PHEME.

	F	C	O	G	S	F1
Our model w/o stance	.228	.267	.300	.333	.293	.405
Our model with stance	.208	.341	.313	.403	.358	.434
SAVED (Dougrez-Lewis et al., 2021)	.372	.351	.304	.281	.332	.434

Table 1: PHEME results for each fold and overall reported as macro-averaged F1 scores. The fold abbreviations stand for Ferguson, Charlie Hebdo, Ottawa Shooting, Germanwings Crash and Sydney Siege

Structure-Aware Model Structure-aware models such as tree-based and graph-based are among the best performing for rumour verification (Kochkina et al., 2018; Bian et al., 2020; Kochkina et al., 2023), given that the task heavily relies on user interactions for determining veracity. Our approach models the conversation thread as a graph, where interactions between posts manifest as propagation (top-down) and dispersion (bottom-up) flows similar to Bian et al. (2020). The architecture contains GraphSage (Hamilton et al., 2017) layers, proven to yield meaningful node representations, followed by GAT (Veličković et al., 2018) layers, which are shown to improve performance in similar tasks (Kotonya et al., 2021; Zhang et al., 2021a; Jia et al., 2022). Sentence Transformers embeddings (Reimers and Gurevych, 2019) are used to initialise the node representations in the graphs. The propagation and dispersion component outputs are each concatenated with the output of a stance component and pooled, resulting in another concatenated representation to which a final multi-head attention layer (Vaswani et al., 2017) is applied.

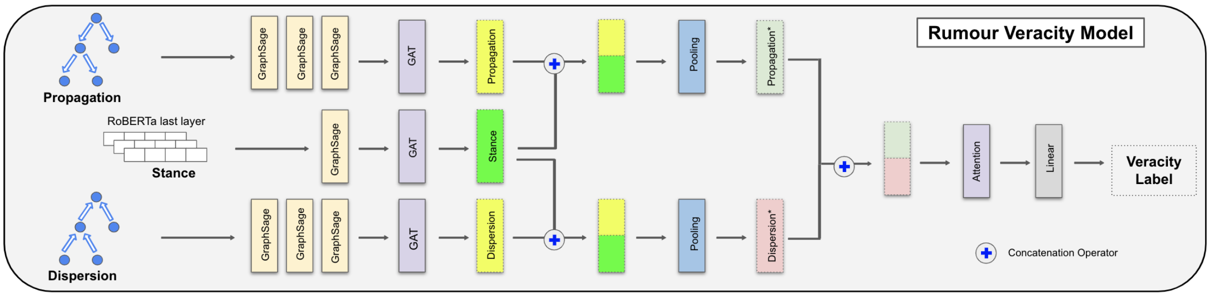

Figure 2: Architecture of our rumour verification model enhanced with structure-aware and stance-aware components based on graph neural networks. In the diagram, Propagation/Dispersion/Dispersion represent the outputs of each respective component, while Propagation*/Dispersion* represent the stance-enriched outputs of these.

Stance-Aware Component Stance detection is closely linked to misinformation detection (Hardalov et al., 2022) with previous work having shown that a joint approach improves rumour verification (Zubiaga et al., 2016; Derczynski et al., 2017; Gorrell et al., 2019; Yu et al., 2020; Dougrez-Lewis et al., 2021). As such our model includes a stance component unlike the GNN by Bian et al. (2020). Since only a small portion of the PHEME dataset is annotated with gold stance labels for the RumourEval competition (Derczynski et al., 2017), we generate silver labels for the whole corpus. In particular, we train a RoBERTa model (Liu et al., 2019) for stance classification and extract the embeddings from the last hidden layer to augment the rumour verification task with stance information. See Appendix D for experimental setup.

Performance of Rumour Verification Baselines We include the performance of our proposed baselines, the structure-aware model and its stance-aware version, in Table 1.

As expected, integrating stance knowledge into the model boosts performance by almost 3 F1-points overall with improved scores across the majority of folds; we hypothesise performance does not improve for the Ferguson fold due to its severe label imbalance skewed towards unverified rumours. Moreover, we observe that the model enhanced with the stance-aware component achieves competitive results and is comparable to the current state-of-the-art model on the PHEME dataset, the SAVED model by Dougrez-Lewis et al. (2021).

3.2 Explaining the Model

3.2.1 Attribution Method

We experiment with two classes of attribution methods: gradient-based and game-theory-based. For gradient-based methods, we choose Integrated Gradients (IG) (Sundararajan et al., 2017). This is a local explainability algorithm that calculates attribution scores for each input unit by accumulating gradients along the interpolated path between a local point and a starting point with no information (baseline). IG was selected over other gradient-based saliency methods such as DeepLIFT (Shrikumar et al., 2017) as it has been shown to be more robust (Pruthi et al., 2022) when applied in classification tasks. Shapley Values (SV) (Štrumbelj and Kononenko, 2014) is the representative explainability method derived from game theory and has been used in many applications (Zhang et al., 2020; Mosca et al., 2021; Mamta and Ekbal, 2022). Its attribution scores are calculated as expected marginal contributions where each feature is viewed as a 'player' within a coalitional game setting.

Note that we focus on post-hoc methods instead of intrinsic ones, such as attention, in our architecture to keep the framework generalisable to other rumour verification models. Specifically, we use IG and SV[2] as methods for $\mathcal{E}$ to calculate the post importance f in Equation 1. This importance with respect to model prediction is then leveraged to sort the posts within the thread in descending order. We then construct subsets of important posts $\mathcal{I}_k \subset \mathcal{I}$ such that $|\mathcal{I}_k| = k\%|\mathcal{I}|$ with $\mathcal{I}_k$ representing the $k\%$ most important posts of the rumour thread, $k = 25, 50, 100$. These will be used as inputs for summarisation in the next stage to determine the trade-off between post importance and number of posts necessary to construct a viable explanation.

3.2.2 Summarisation for Explanation

We propose explanation baselines spanning different generation strategies: extractive vs abstractive, model-centric vs model-independent and in-domain vs out-of-domain.

[2]Used *captum* package (Kokhlikyan et al., 2020) for both.

Extractive Explanations

- *Important Response*: the response within the thread scored as most important by each attribution method. This is a model-dependent explanation.
- *Similar Response*: the response within the thread most semantically similar to the source claim, as scored by SBERT (Reimers and Gurevych, 2019). This model-independent baseline is inspired by (Russo et al., 2023).

Abstractive explanations have a dual purpose that fits the challenging set-up of our pipeline: they serve as a way to aggregate important parts of the thread, and also provide a fluent justification sourced from multiple views to a claim's veracity.

- *Summary of $\mathcal{I}$*: We summarise the set $\mathcal{I}$ of important posts to obtain a model-centric explanation. We fine-tune BART (Lewis et al., 2020) on the MOS corpus introduced by Bilal et al. (2022) that addresses summarisation of topical groups of tweets by prioritising the majority opinion expressed. We hypothesise this template-guided[3] approach will satisfy the explanatory purpose since user opinion is an important indicator for assessing a claim's veracity in rumour verification (Hardalov et al., 2022). Similarly, we define explanations *Summary of* $\mathcal{I}_{25}$ & *Summary of* $\mathcal{I}_{50}$.
- *Out-of-domain Summary*: We use the BART (Lewis et al., 2020) pre-trained on the CNN/Daily Mail (Nallapati et al., 2016) dataset without any fine-tuning and summarise the entire thread. This yields a model-independent explanation.

We note that while supervised summarisation is used to inform our generation strategy, our resulting explanations never rely on gold explanations annotated for the downstream task of fact-checking. In fact, neither MOS (Bilal et al., 2022) nor the CNN/Daily Mail (Nallapati et al., 2016) datasets were aimed for fact-checking and both focus on broad topics unrelated to the PHEME claims.

4 Automatic Evaluation of Explanation Quality

As the PHEME dataset lacks gold standard explanations to compare against, we prioritise the extrinsic evaluation of the generated explanations with respect to their usefulness in downstream tasks. This is similar to work on explanatory fact-checking (Stammbach and Ash, 2020; Krishna et al., 2022).

[3]The template summary takes the form: *Main Story* + *Majority* Opinion expressed in the thread.

You will be shown a **Claim** and an **Explanation**. The veracity of the Claim can either be true, false or unverified. Choose an option from A to D that answers whether the Explanation can help confirm the veracity of the Claim.

A: The Explanation confirms the information in the Claim is true. The Explanation will include evidence to prove the Claim or show users believing the Claim.

B: The Explanation confirms the information in the Claim is false. The Explanation will include evidence to disprove the Claim or show users denying the Claim.

C: The Explanation confirms the information in the Claim is unverified. The Explanation will state no evidence exists to prove or disprove the Claim or show users doubting the Claim.

D: The Explanation is irrelevant in confirming the veracity of the Claim. The Explanation will not include any mention of evidence and users will not address the veracity of the Claim.

Claim: {claim}
Explanation: {explanation}

Table 2: Example task instructions used in the prompt following a multiple-choice setting.

In particular, we use the criterion of **informativeness** defined by Atanasova et al. (2020) as the ability to deduce the veracity of a claim based on the explanation. If the provided explanation is not indicative of any veracity label (*true*, *false*, or *unverified*), the explanation is considered uninformative. Otherwise, we compare the veracity suggested by the explanation to the prediction made by the model. This enables the evaluation of the explanation's fidelity to the model and is one of the main approaches to assess explanatory **faithfulness** in the research community (Jacovi and Goldberg, 2020).

We devise a novel evaluation strategy for capturing the informativeness of generated explanations based on LLMs. This is motivated by recent work demonstrating the effectiveness of LLM reasoning capability in various tasks (Kojima et al., 2022; Chen, 2023), including as a zero-shot evaluator for summarisation outputs (Liu et al., 2023; Shen et al., 2023; Wang et al., 2023a). We use OpenAI's *gpt-3.5-turbo-0301*[4], hereinafter referred to as ChatGPT, which is a snapshot of the model from 1 March 2023 that will not receive updates – this should encourage the reproducibility of our evaluation. We follow a multiple-choice setting in the prompt similar to Shen et al. (2023). Our initial experiments confirmed previous findings (Brown et al., 2020) that GPT reasoning can be improved by including a few annotated representative examples of the evaluation within its prompt (See Appendix A). We experimented with several prompt designs varying in level of detail (no justification of answer, no examples) and found that the most exhaustive prompt yielded best results. The final task instructions used for the prompt are in Table 2.

[4]Used GPT-3.5-turbo due to its lower running costs compared to GPT-4.

We ran a pilot study (See Appendix C) to establish which temperature setting yields the most robust LLM evaluation. To account for any non-deterministic behaviour, the experiment was run three times. We find the results remain 100% consistent across runs for temperature 0. As this is in line with the settings used in similar works employing LLMs as evaluators (Shen et al., 2023), we also use this value for our experiment. Each request is sent independently via the Open AI API. Since using an LLM evaluator allows us to scale our evaluation (Chiang and Lee, 2023), we use all suitable PHEME threads[5] (i.e. 1233 / 2107 threads) for testing. This set-up foregoes the costs necessary to obtain a diverse manually-annotated test set and offers more statistical power to the results as recommended by Bowman and Dahl (2021).

5 Results and Discussion

The results are shown in Table 3.

	Uninformative	Unfaithful	Faithful
	Extractive Explanations		
Important Response (IG)	67.23	**21.33**	11.44
Important Response (SV)	65.29	22.30	12.41
Similar Response	30.98	43.88	25.14
	Abstractive Explanations		
Summary of $\mathcal{I}_{25}$ (IG)	23.68	46.55	29.76
Summary of $\mathcal{I}_{25}$ (SV)	22.95	48.50	28.55
Summary of $\mathcal{I}_{50}$ (IG)	**22.11**	46.47	**30.41**
Summary of $\mathcal{I}_{50}$ (SV)	23.60	47.20	29.20
Summary of $\mathcal{I}$ (IG)	24.90	48.58	26.52
Summary of $\mathcal{I}$ (SV)	23.60	48.90	27.49
Out-of-domain Summary	39.17	38.28	22.55

Table 3: Explanation evaluation wrt model prediction (%). If the explanation cannot be used to infer a veracity label for the claim, it is **uninformative**. Otherwise, the explanation is **faithful** if its label coincides with the prediction and **unfaithful** if not. Best scores are in bold.

Model-centric vs Model-independent We note that the explanations *Out-of-domain Summary* and *Similar Response* are independent of the rumour verification model built in section 3.1 as they are not produced by any of the post-hoc algorithms. Hence, while these are not expected to be faithful, we analyse how they compare in informativeness to the other model-centric explanations. We find that abstractive explanations (*Summaries of* $\mathcal{I}_{25}$, $\mathcal{I}_{50}$, $\mathcal{I}$) informed by the rumour verifier are the most informative of all. Thus, summarising a selection of important posts learned during the rumour verification process yields a better explanation than relying on individual replies or summarising the whole thread.

Integrated Gradients vs Shapley Values The summaries generated via IG achieve better scores than the SV ones in both informativeness and faithfulness. While SV initially provides a better *Important Response*, it fails to detect other important posts within the thread as suggested by the scores for I_{25} and I_{50}. Moreover, the time complexity for the SV algorithm is exponential as its sampling strategy increases proportionately with the number of perturbed input permutations. We note the average computation time for both algorithms to assess a thread of 15 posts: 0.5s for IG and 2011s for SV. This makes IG a more suitable algorithm with respect to both performance and running time.

Extractive Explanation The best extractive baseline is the *Similar Response*, which selects the closest semantic match from the thread to the claim. Followed by are model-centric baselines *Important Response* for both IG and SV, lagging behind by a large margin. We investigate the reason behind this performance by checking the stance labels of the corresponding posts. Using the labelled data from Derczynski et al. (2017), we train a binary RoBERTa to identify comments and non-comments[6] where a comment is defined as a post that is unrelated or does not contribute to a rumour's veracity. We find that 64% of posts corresponding to *Important Response* labelled as uninformative are also classified as comments, much higher than 47% for *Similar Response*. This explains why semantic similarity can uncover a more relevant explanation than the *Important Response* alone. Still, this method suffers from 'echoing' the claim [7], which risks missing out on other important information found in the thread (see Table 4).

Abstractive Explanation The abstractive explanations are shown to be considerably more informative than most extractive baselines. They have the advantage of aggregating useful information that appears later in the conversation. For instance, the abstractive explanations in Table 4 indicate posters' doubt and requests for more details. Furthermore, using an opinion-driven summariser is better for constructing a more informa-

[5]Suitable defined as at least ten posts and the majority are non-empty after URL and user mentions are removed.

[6]The original task is a 4-way classification of posts into one of the stance labels: *support*, *deny*, *query*, or *comment*. This is simplified by aggregating the first three labels into one.

[7]The majority of informative *Similar Responses* are classified as supporting the claim.

Claim
Update from Ottawa: Cdn soldier dies from shooting -Parliamentary guard wounded Parliament Hill still in lockdown URL
Prediction: Unverified
Explanation Summaries
Important Response: @TorontoStar Ok, time to take it to the *** muslims. Look out Allah, here comes the revenge. ***. (*Uninformative*) **Similar Response**: #AttackinOttawa @TorontoStar: Update Cdn soldier dies from shooting -Parliamentary guard wounded Parliament Hill still in lockdown URL (*True*)
Summary of $\mathcal{I}_{25}$ (IG): Soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. The majority think the media is wrong to report that Parliament Hill was in lockdown and that the lockdown was a ploy to target Muslims. (*False*)
Summary of $\mathcal{I}_{50}$ (IG): Cdn soldier dies from shooting dead in Ottawa. The majority are sceptical about the news of the shooting and some are questioning where the confirmation is coming from. (*Unverified*) **Summary of $\mathcal{I}$ (IG)**: Cdn soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. Most users ask where the news of the gunman is and are wondering who is responsible for his death. Many of the responses use humour and irony, such as: 'I don't think the soldier is dead'. (*Unverified*) **Out-of-domain Summary**: Update from Ottawa: -Cdn soldier dies from shooting -Parliamentary guard wounded. It looks like confirmations are coming in now. I don't think the soldier is dead. (*Unverified*)

Table 4: Example explanation summaries. Manually-annotated red highlights explain the model prediction for the given claim. ChatGPT evaluations are in ().

tive summary-explanation than other options (See Sec. 3.2). We have also investigated the degree of information decay in relation to the number of posts used for summary construction in model-centric explanations. In Table 3, the summary based on the first half of important posts ($\mathcal{I}_{50}$) yields the most informative and faithful explanation for both algorithms, closely followed by the $\mathcal{I}_{25}$ one. The worst-performing model-centric explanation is that generated from the whole set of important replies ($\mathcal{I}$). We calculate the cumulative importance score of these data partitions and note $\mathcal{I}_{25}$ and $\mathcal{I}_{50}$ contain 75% and 93% respectively of the thread's total importance. This suggests the remaining second half of the importance-ordered thread offers little relevant information towards the model's decision.

6 Human Evaluation of LLM-based Evaluators

Agreement	Informativeness Detection	Veracity Prediction
Ann - Ann	82%	88%
Ann - ChatGPT	69%	68%
Ann - ChatGPT 0613	64%	74%
Ann - GPT-4	63%	80%

Table 5: Pairwise agreement scores for the overlap between the evaluations of the annotators (Ann) and the LLM. The LLMs are: ChatGPT ("gpt-3.5-turbo-0301"), ChatGPT 0613 ("gpt-3.5-turbo-0613") and GPT-4. The evaluations are conducted for two tasks: informativeness detection and veracity prediction.

Our human evaluation study has two goals: 1) quantify the evaluation capability of ChatGPT, the LLM employed in our experiments in Sec. 5 to assess automatic explanations and 2) investigate the performance of ChatGPT against more recently-published LLMs. The results are in Table 5.

We ran a pilot study on 50 threads randomly sampled, such that each fold and each label type is equally represented for a fair evaluation of the LLM performance. We follow a similar evaluation setup to the work of (Atanasova et al., 2020), who study whether their generated summaries provide support to the user in fact checking a claim. We check the LLM-based evaluation of automatic explanations on two tasks: 1. **Informativeness Detection**, where an Explanation is classified as either informative or uninformative and 2. **Veracity Prediction**, where an Informative Explanation is assigned true, false or unverified if it helps determine the veracity of the given claim.

Two Computer Science PhD candidates proficient in English were recruited as annotators for both tasks. Each annotator evaluated the test set of explanation candidates, resulting in 300 evaluations per annotator. The same guidelines included in the prompt from Table 2 and examples from Appendix A are used as instructions. Before starting, the research team met with the annotators to ensure the tasks were understood, a process which lends itself to a richer engagement with the guidelines.

6.1 Evaluation of ChatGPT

Informativeness Detection In our first human experiment (Table 5: first column), we evaluate whether ChatGPT correctly identifies an informative explanation. We find that the agreement between our annotators is 82% which we set as the upper threshold for comparison. We note that the agreement between human evaluators and ChatGPT consistently remains above the random baseline, but experiences a drop. Fleiss Kappa is $\kappa = 0.441$, which is higher than the agreement of $\kappa = 0.269, 0.345, 0.399$ reported by Atanasova et al. (2020) for the same binary setup. After examining the confusion matrix for this task (See Appendix B), it is observed that most mismatches arise from false positives - ChatGPT labels an Explanation as informative when it is not. Finally, we find this type of disagreement occurs in instances when the rumour is a complex claim, i.e., a claim with more than one check-worthy piece of information within it. As suggested by Chen et al. (2022), the analysis of complex real-world claims is a challenging task in the field of fact checking and we also observe its impact on our LLM-based evaluation for rumour verification.

Veracity Prediction In our second human experiment (Table 5: second column), we evaluate if ChatGPT correctly assigns a veracity label to an Informative explanation. Again, we consider 88%, the task annotator agreement to be the upper threshold. Despite the more challenging set-up (ternary classification instead of binary), the LLM maintains good agreement: Fleiss Kappa $\kappa = 0.451$ (again higher than those of Atanasova et al. (2020) for the multi-class setup $\kappa = 0.200, 0.230, 0.333$). Manual inspection of the disagreement cases reveals that the most frequent error type (58 / 75 mislabelled cases exhibit this pattern - See Appendix B) is when ChatGPT classifies a rumour as unverified based on the Explanation, while the annotator marks it as true. We hypothesise that an LLM fails to pick up on subtle cues present in the explanation that are otherwise helpful for deriving a veracity assessment. For instance, the Explanation *"I think channel 7 news is saying he [the hostage-taker] is getting agitated bcoz of it [the hostage's escape], its time to go in."* implies that the escape indeed took place as validated by Channel 7; this cue helps the annotator assign a true label to the corresponding claim *"A sixth hostage has escaped from the Lindt cafe in Sydney!"*.

We acknowledge the limitations of using an LLM as an evaluator, which reduces the richness of annotator interaction with the task, but show through our human evaluations that good agreement between an LLM and humans can still be achieved. This not only allows the scaling of final results to the entire dataset instead of being confined to a small test set (See Sec. 4), but also provides an automated benchmarking of generated explanations when the ground truth is missing.

6.2 Comparison to other LLMs

As ChatGPT is a closed-source tool continually updated by its team, it is important to investigate how ChatGPT-powered evaluations are influenced by the release of newer versions of the same language model or by substitution with improved models. To this effect, we compare the legacy version of ChatGPT released on 1 March 2023 with its more recent version, ChatGPT 0613 (released on 13 June 2023) and finally with GPT-4, a multimodal model equipped with broader general knowledge and more advanced reasoning capabilities.

We note that that while there are differences between the labels produced by the two versions, there is a higher agreement with human judgement for the newer snapshot ChatGPT 0613 when assessed on the more complex task of veracity prediction. A similar behaviour is observed for GPT-4, whose performance is the most aligned with human judgment in the second task. After examining the error patterns (See Appendix B), we observe a notable difference between ChatGPT-based models and GPT-4: while both temporal snapshots of ChatGPT tend to evaluate irrelevant explanations as informative (See Sec. 6.1), GPT-4 suffers from assigning too many false negatives. This implies the existence of a positive bias for ChatGPT models and a negative bias for GPT-4.

Based on our limited findings, we hypothesise that more recent models have the potential to be more reliable evaluators of explanations than older models, given their higher agreement with human annotators. However, the model choice needs to be grounded into the task requirements (i.e., which errors should be prioritised) and availability of computational costs (at the moment of writing GPT-4 is 20x more expensive than ChatGPT).

7 Conclusions and Future Work

We presented a novel zero-shot approach for generating abstractive explanations of model predictions for rumour verification. Our results showed abstractive summaries constructed from important posts scored by a post-hoc explainer algorithm can be successfully used to derive a veracity prediction given a claim and significantly outperform extractive and model-independent baselines. We also found using an LLM-based evaluator for assessing the quality of the generated summaries yields good agreement with human annotators for the tasks of informativeness detection and veracity prediction.

In future work, we plan to jointly train the veracity prediction and explanation generation and assess how an end-to-end approach impacts the quality of resulting explanations. Additionally, we aim to enrich the explanations by incorporating external sources of information such as PHEMEPlus (Dougrez-Lewis et al., 2022). Another direction is generating fine-grained explanations for addressing all check-worthy aspects within complex claims.

Limitations

Summarisation of threads The format of the conversation threads is challenging to summarise. Our approach to summarisation is to flatten the conversation tree and to concatenate the individual posts, which are then used as an input to a BART model. This approach is naïve as the meaning of the nested replies can be lost if considered independently of the context.

Task limitation At the moment, the explanations are constructed exclusively from the information present in the thread. Consequently, the degree of evidence present in a thread is reflected into the explanatory quality of the summary.

Complex Claims As seen in the paper, complex claims are a challenging subset of rumours to evaluate. Using the heuristic outlined in Chen et al. (2022) to identify complex claims based on verb count, we find that 22% of the claims within PHEME are classified as complex. To generate comprehensive explanations covering each check-worthy aspect within such claims, a re-annotation of PHEME is required which is only labelled at claim-level at the moment.

Human Evaluation Evaluation via large language models is in its infancy. While there have been very encouraging recent results of using it as a viable alternative to human evaluation, these are still early days. It is unclear how much the evaluation stability is impacted by prompt design or by substitution with open-source language models.

Evaluation criteria for generated output Since our explanations rely on generation mechanisms including automatic summarisers, it is important to acknowledge that there are other evaluation criteria native to the generation field which are outside the scope of this paper and have not been covered. We note that since hallucination, redundancy, coherence and fluency have already been tested in the original works (Lewis et al., 2020; Bilal et al., 2022) introducing the summarisers we employ, we prioritised the criteria relevant to explainable fact-checking in the experiments of this paper: informativeness of explanations and faithfulness to predicted veracity label.

Ethics Statement

Our experiments use PHEME dataset, was given ethics approval upon its original release. However, we note that the dataset contains many instances of hate speech that may corrupt the intended aim of the summaries. In particular, summaries that use the majority of posts within the thread may exhibit hate-speech content exhibited by parts of the input text.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. *Conference for Truth and Trust Online*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. 2022. Template-based abstractive microblog opinion summarization. *Transactions of the Association for Computational Linguistics*, 10:1229–1248.

Alodie Boissonnet, Marzieh Saeidi, Vassilis Plachouras, and Andreas Vlachos. 2022. Explainable assessment of healthcare articles with QA. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. 2021. Learning disentangled latent topics for Twitter rumour veracity classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3902–3908, Online. Association for Computational Linguistics.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation?

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Hao Jia, Honglei Wang, and Xiaoping Zhang. 2022. Early detection of rumors based on source tweet-word graph attention networks. *PLoS One*, 17(7):e0271224.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10).

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and abstractive explanations for fact-checking and evaluation of news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–50, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Elena Kochkina, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing and Management*, 60(1):103116.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021. Graph reasoning with context-aware linearization for interpretable fact extraction and verification. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 21–30, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

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. *CoRR*, abs/1907.11692.

. Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 478–492, Online only. Association for Computational Linguistics.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Rob Procter, Miguel Arana-Catania, Yulan He, Maria Liakata, Arkaitz Zubiaga, Elena Kochkina, and Runcong Zhao. 2023. Some observations on fact-checking work with implications for computational support.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization?

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries forautomated fact checking. In *Conference for Truth and Trust Online*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations*.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023. Less is more for long document summary evaluation by llms.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401, Online. Association for Computational Linguistics.

Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang. 2020. An explainable machine learning framework for fake financial news detection. In *Proceedings of the 41st International Conference on Information Systems, ICIS 2020, Making Digital Inclusive: Blending the Locak and the Global, Hyderabad, India, December 13-16, 2020*. Association for Information Systems.

Xinpeng Zhang, Shuzhi Gong, and Richard O. Sinnott. 2021a. Social media rumour detection through graph attention networks. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6.

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021b. Faxplainac: A fact-checking tool based on explainable models with human correction in the loop. CIKM '21, page 4823–4827, New York, NY, USA. Association for Computing Machinery.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11:1–29.

A Examples of Assessing the Informativeness of Explanations

Claim: Victims were forced to hold a flag on the cafe window.
Explanation: Users believe this is true and point to the released footage.
Your answer: A

Claim: BREAKING: Hostages are running out of the cafe #sydneysiege
Explanation: Some users believe the claim is unverified as Channel 9 did not confirm and some agree that the details of potential escape should not be disclosed.
Your answer: C

Claim: One of the gunmen left an ID behind in the car.
Explanation: One of the gunmen left an ID behind in the car. The majority deny the ID was found there and point to the media for blame.
Your answer: B

Claim: Three people have died in the shooting.
Explanation: Three people have died in the shooting. Most users pray the attack is over soon.
Your answer: D

Claim: NEWS #Germanwings co-pilot Andreas Lubitz had serious depressive episode (Bild newspaper) #4U9525 URL LINK
Explanation:Germanwings co-pilot Andrés Lubitz has serious depressive episode. Never trust bild. Users believe that bild is a fake newspaper and the stories concerned with the suicide of Andreas Lubitz should not be discussed.
Your answer: C

Claim: Snipers set up on National Art Gallery as we remain barricaded in Centre Block on Parliament Hill #cdnpoli.
Explanation: Snipers set up on National Art Gallery as we remain barricaded in Centre Block on Parliament Hill. Most users are skeptical about the news and await more details.
Your answer: C

Claim: BREAKING: #Germanwings co-pilot's name is Andreas Lubitz, a German national, says Marseilles prosecutor.
Explanation: He didn't have a political or religious background.
Your answer: D

Claim: Several bombs have been placed in the city
Explanation: This is false, why then cause panic and circulate on social media?
Your answer: B

Claim: Police report the threats released by the criminals.
Explanation: The majority threaten to condemn anyone who is a terrorist.
Your answer: D

Claim: #CharlieHebdo attackers shouted 'The Prophet is avenged'.
Explanation: In video showing assassination of officer.walking back to car they shouted: 'we avenged the prophet.We killed Charlie Hebdo'
Your answer: A

Table 6: Ten representative examples covering diverse explanation styles and veracity labels are selected. These are included in the final prompt for ChatGPT.

B Error Analysis of LLM's performance as Evaluator

We note that our ChatGPT-human agreement scores for both tasks are similar or higher to those reported by Zubiaga et al. (2016), who employ crowd-sourced workers for annotating similar classification subtasks on the PHEME dataset: 61.1% for labelling certainty of rumours and 60.8% for classifying types of evidence arising from the thread.

We report the performance of ChatGPT, ChatGPT 0614 and GPT-4 as evaluators using the manually annotated set of 200 explanations. The error analysis is shared via a confusion matrix for each task: informativeness detection (See Table 7) and veracity prediction (See Table 8). The results are reported as counts.

C Pilot Study on Temperature Setting for ChatGPT

We used the same explanations in Table 4 and ran a small pilot study to assess how incrementing the temperature parameter affects the LLM evaluation. Results are in Table 9. We used increments of 0.2 in temperature and ran the experiment 3 times to account for the non-deterministic behaviour. Overall, the evaluations remain consistent (94% of the labels output by ChatGPT are the same) across runs

LLM \ Annotator	Informative	Uninformative
	ChatGPT	
Informative	169	107
Uninformative	81	143
	ChatGPT 0613	
Informative	236	104
Uninformative	114	146
	GPT-4	
Informative	160	30
Uninformative	190	220

Table 7: Confusion Matrices for ChatGPT, ChatGPT 0613 and ChatGPT-4 for the task of **Informativeness Detection**

LLM \ Annotator	True	False	Unverified
		ChatGPT	
True	105	3	4
False	12	18	5
Unverified	58	3	61
		ChatGPT 0613	
True	114	3	8
False	10	10	6
Unverified	26	8	51
		GPT-4	
True	78	0	2
False	10	10	9
Unverified	7	84	40

Table 8: Confusion Matrices for ChatGPT, ChatGPT 0613 and ChatGPT-4 for the task of **Veracity Prediction**

and temperature values. In particular, we note that when using temperature 0, the evaluations remain 100% consistent and for non-zero temperature, the evaluation only impacts the labelling of the last explanation which is less helpful than previous explanation candidates.

D Experimental Setup

We train the rumour verification model for 300 epochs with learning rate 10^{-5}. The training loss is cross-entropy. The optimizer algorithm is Adam (Kingma and Ba, 2015). Hidden channel size is set as 256 for the propagation and dispersion components and 32 hidden channel size for the stance component. The batch size is 20. For the GraphSage layers, we apply a mean aggreggator scheme, followed by a relu activation. For the Multi-headed Attention layer, we use 8 heads. Embeddings generated by the "all-MiniLM-L6-v2" model from Sentence Transformers (Reimers and Gurevych, 2019) are used to initialise the node representations in the graphs. To avoid overfitting, we randomly dropout an edge in the graph networks with probability 0.1. We use a Nvidia A5000 GPU for our model training. All model implementation is done via the *pytorch-geometric* package (Fey and Lenssen, 2019) for graph neural networks.

Explanation	$T = 0$	$T = 0.2$	$T = 0.4$	$T = 0.6$	$T = 0.8$	$T = 1$
@TorontoStar Ok, time to take it to the ***muslims. Look out Allah, here comes the revenge. ***.	D,D,D	D,D,D	D,D,D	D,D,D	D,D,D	D,D,D
Soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. The majority think the media is wrong to report that Parliament Hill was in lockdown and that the lockdown was a ploy to target Muslims.	B,B,B	B,B,B	B,B,B	B,B,B	B,B,B	B,B,B
Cdn soldier dies from shooting dead in Ottawa. The majority are sceptical about the news of the shooting and some are questioning where the confirmation is coming from.	C,C,C	C,C,C	C,C,C	C,C,C	C,C,C	C,C,C
Cdn soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. Most users ask where the news of the gunman is and are wondering who is responsible for his death. Many of the responses use humour and irony, such as: 'I don't think the soldier is dead'.	C,C,C	C,A,C	C,C,C	C,C,C	C,A,A	C,C,A

Table 9: Labels output by ChatGPT for each explanations across 3 different runs.

E Current Submission colour-coded for the changes we have implemented compared to the previous version of the manuscript

Red stands for removed material and blue stands for new additions.

Generating ~~Unsupervised~~ Zero-shot Abstractive Explanations for Rumour Verification

Anonymous ACL submission

Abstract

The task of rumour verification in social media concerns assessing the veracity of a claim on the basis of conversation threads that result from it. While previous work has focused on predicting a veracity label, here we reformulate the task to generate model-centric, free-text explanations of a rumour's veracity. The approach is model agnostic in that it generalises to any model. Here we propose a novel GNN-based rumour verification model. follow ~~an unsupervised~~ a zero-shot approach by first ~~utilising~~ applying post-hoc explainability methods to score the most important posts within a thread and then we use these posts to generate informative explanatory summaries ~~by employing template-guided~~ using opinion-guided summarisation. To evaluate the informativeness of the explanatory summaries, we exploit the few-shot learning capabilities of a large language model (LLM). Our experiments show that LLMs can have similar agreement to humans in evaluating summaries. Importantly, we show that explanatory abstractive summaries are more informative and better reflect the predicted rumour veracity than just using the highest ranking posts in the thread.[1]

1 Introduction

Evaluating misinformation on social media is a challenging task that requires many steps (Zubiaga et al., 2016): detection of rumourous claims, identification of stance towards a rumour, and finally assessing rumour veracity. In particular, misinformation may not be immediately verifiable using reliable sources of information such as news articles since they might not have been available at the time a rumour has emerged. For the past eight years, researchers have focused on the task of automating the process of rumour verification in terms of assigning a label of *true*, *false*, or *unverified* (Zubiaga et al., 2016; Derczynski et al., 2017). However, recent work has shown that while fact-checkers agree with the urgent need for computational tools for content verification, the output of the latter can only be trusted if it is accompanied by explanations (Procter et al., 2023).

Thus, in this paper, we move away from black-box classifiers of rumour veracity to generating explanations written in natural language (free-text) for why, given some evidence, a statement can be assigned a particular veracity status. This has real-world applicability particularly in rapidly evolving situations such as natural disasters or terror attacks (Procter et al., 2013), where the explanation for an automated veracity decision is crucial (Lipton, 2018). To this effect, we use a popular benchmark, the PHEME (Zubiaga et al., 2016) dataset, to train a rumour ~~verification model~~ verifier and employ the conversation threads that form its input to generate model-centric explanation summaries of the model's assessments.

Atanasova et al. (2020), Kotonya and Toni (2020) and Stammbach and Ash (2020) were the first to introduce explanation summaries for fact-checking across different datasets. Kotonya and Toni (2020) ~~provide~~ provided a framework for creating abstractive summaries that justify the true veracity of the claim in the PUBHealth dataset, similarly to Stammbach and Ash (2020) who augment the FEVER (Thorne et al., 2018) dataset with a corpus of explanations. Atanasova et al. (2020) proposed a jointly trained system that produces veracity predictions as well as explanations in the form of extracted evidence from ruling comments on the LIAR-PLUS dataset (Alhindi et al., 2018). The approach in (Kotonya and Toni, 2020) results in explanatory summaries that are~~however~~, however, not faithful to the model, while Atanasova et al. (2020) requires a super-

[1]~~We supplement a~~ A sample of ~~our~~ generated explanations and ~~our source~~ code ~~which we will fully release on a GitHub repository after the anonymity period~~are provided.

vised approach. Our goal is to create a novel ~~unsupervised~~ zero-shot method for abstractive explanations that ~~are faithful to~~ explain the rumour verification model's predictions. We make the following contributions:

- We introduce ~~an unsupervised~~ a zero-shot framework for generating abstractive explanations using ~~template-guided~~ opinion-guided summarisation for the task of rumour verification. To the best of our knowledge, this is the first time ~~that~~ free-text explanations are introduced for this task.
- We investigate the benefits of using a gradient-based algorithm and a game theoretical algorithm to provide explainability. ~~to a novel graph-based hierarchical model for rumour verification.~~
- While our explanation generation method is generalisable to any verification model, we introduce a novel graph-based hierarchical approach.
- We evaluate the informativeness of several explanation baselines, including model-independent and model-dependent ones stemming from the highest scoring posts by providing them as input to a few-shot trained large language model. Our results show that ~~abstractive explanations are informative in 75%~~ our proposed abstractive model-centric explanations are more informative in 77% of the cases as opposed to ~~34% for the highest ranked post~~49% for all other baselines.
- We provide both human and LLM-based evaluation of the generated ~~explanatory summaries~~ explanations, showing that LLMs achieve sufficient agreement with humans, thus allowing ~~to scale~~ scaling of the evaluation of the explanatory summaries in absence of gold-truth explanations.

2 Related Work

Explainable Fact Checking Following the example of fact-checking ~~platforms~~ organisations (e.g., Snopes, Full Fact, Politifact), which provide journalist-written justifications to determine the truthfulness of claims, recent datasets augmented with free-text explanations have been constructed: LIAR-PLUS (Alhindi et al., 2018), PubHealth (Kotonya and Toni, 2020), AVeriTeC (Schlichtkrull et al., 2023). A wide range of explainable outputs and methods have been proposed: theorem proofs (Krishna et al., 2022), knowledge graphs (Ahmadi et al., 2019), question-answer decompositions (Boissonnet et al., 2022; Chen et al., 2022), reasoning programs (Pan et al., 2023), deployable evidence-based tools (Zhang et al., 2021b) and summarisation (Atanasova et al., 2020; Kotonya et al., 2021; Stammbach and Ash, 2020; Kazemi et al., 2021; Jolly et al., 2022). We adopt summarisation as our generation strategy as it fluently aggregates evidence from multiple inputs and has been proven effective in similar works which we discuss next.

~~We discuss the work on summarisation in more detail in the next paragraph.~~

Explainability as Summarisation Atanasova et al. (2020) and Kotonya and Toni (2020) leveraged large-scale datasets annotated with gold justifications to generate supervised explanations for fact-checking, while Stammbach and Ash (2020) used few-shot learning on GPT-3 to create the e-FEVER dataset of explanations. Similar to (Stammbach and Ash, 2020), Kazemi et al. (2021) also leveraged a GPT-based model (GPT-2) to generate abstractive explanations, but found that that their extractive baseline, Biased TextRank, outperformed GPT-2 on the LIAR-PLUS dataset (Alhindi et al., 2018). Jolly et al. (2022) warn that the output of extractive explainers lacks fluency and sentential coherence, which motivated their work on unsupervised post-editing using the explanations produced by Atanasova et al. (2020). Our approach is different from the above as we derive our summaries from microblog content (as opposed to news articles as done by Atanasova et al. (2020); Stammbach and Ash (2020); Kazemi et al. (2021); Jolly et al. (2022), and only use the subset of posts relevant to the model's decision to inform the summary (rather than summarising the whole input as in (Kotonya and Toni, 2020; Kazemi et al., 2021). Moreover, we rely on ~~an unsupervised~~ a zero-shot generation approach without gold explanations, contrary to (Atanasova et al., 2020; Kotonya and Toni, 2020).

LLMs as evaluators Having generated explanatory summaries the question arises as to how to evaluate them at scale. Large language models have been employed as knowledge bases for fact-checking (Lee et al., 2020; Pan et al., 2023), as explanation generators for assessing a claim's veracity (Stammbach and Ash, 2020; Kazemi et al., 2021) and, as of recently, as evaluators in generation tasks. Most works focused on assessing the capability of LLM-

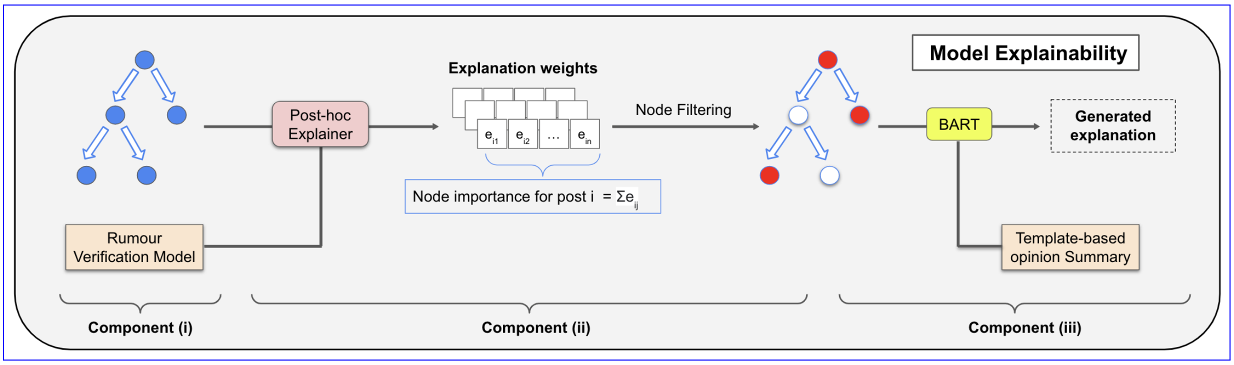

Figure 1: Framework of our proposed approach to obtain faithful generated explanations for the rumour verification model. It explains the process of explanation generation, where the weights from a model are passed through an explainer algorithm to identify important input nodes, which are then filtered and used in abstractive summarisation.

based evaluation on summarisation tasks~~such as long-document summarisation (Wu et al., 2023) and summarisation~~, either on long documents (Wu et al., 2023) or for low-resource languages (Hada et al., 2023). While there is work focusing on reducing positional bias (Wang et al., 2023b) and costs incurred (Wu et al., 2023) for using LLM-based evaluators, our evaluation is most similar to Liu et al. (2023); Shen et al. (2023); Chiang and Lee (2023), who study the extent of LLM-human agreement in evaluations of fine-grained dimensions, such as fluency or consistency. We believe ~~to be~~ we are the first to use an LLM-powered evaluation to assess the informativeness and faithfulness of explanations for verifying a claim.

3 Methodology

Our methodological approach (Figure 1) consists of three individual components: *i)* training a rumour verification model; *ii)* using a post-hoc explainability algorithm; *iii)* generating ~~explanations via abstractive template-based summarisation~~summary-explanations. The approach to explanation generation is ~~unsupervised~~ zero-shot and model-agnostic.

We demonstrate our approach on PHEME (Zubiaga et al., 2016), a widely used benchmark dataset for classifying social media rumours into either unverified, true or false. It contains conversation threads ~~which~~ that cover 5 real-world events such as the Charlie Hebdo attack and the Germanwings plane crash. We adopt the same leave-one-out testing ~~approach~~ as previous works (Dougrez-Lewis et al., 2022) which favours real-world applicability as the model is tested on new events not included in the test data.

Task Formulation For a model trained on rumour verification $\mathcal{M}$, an attribution-based explanation method $\mathcal{E}$, and a rumourous conversation thread consisting of posts $\mathcal{T} = \{p_1, ...p_l\}$ with embeddings $\{x_1, ...x_l\} \subset \mathbb{R}^n$, we define the post importance as a function $f_{(\mathcal{M},\mathcal{E})} : \mathcal{T} \rightarrow \mathbb{R}$.

$$f_{(\mathcal{M},\mathcal{E})}(p_i) = \sum_{j=1}^{n} \mathcal{E}(\mathcal{M}, x_i)_j = \sum_{j=1}^{n} e_{ij} \quad (1)$$

where ~~e_{ij}~~ $e_i \in \mathbb{R}^n$ is the attribution ~~score assigned by the explainer algorithm to the jth position of~~ vector for embedding x_i ~~for~~ of post p_i ~~and n is the size of the embedding vectors.~~ such that each value e_{ij} corresponds to the weight of feature x_{ij} assigned by the explainer algorithm $\mathcal{E}$.

The summary is generated from the subset of posts that are most important for the model prediction, i.e., $\mathcal{I} = \{p_i \mid f_{(\mathcal{M},\mathcal{E})}(p_i) > 0\}$. Note a thread will contain posts that agree with the final prediction (positive importance scores) and posts that disagree (negative importance scores).

3.1 Rumour Verification Model

Our explanation generation method is applicable to any rumour verification model, but here we chose an approach based on graph neural networks (See Figure 2), which caters for a flexible information structure ~~that combines~~ combining information in the conversation thread with information about stance. This is the first time a GNN-based model enriched with stance has been proposed for PHEME.

~~**Structure-Aware Model**~~

Structure-Aware Model Structure-aware models such as tree-based and graph-based are among

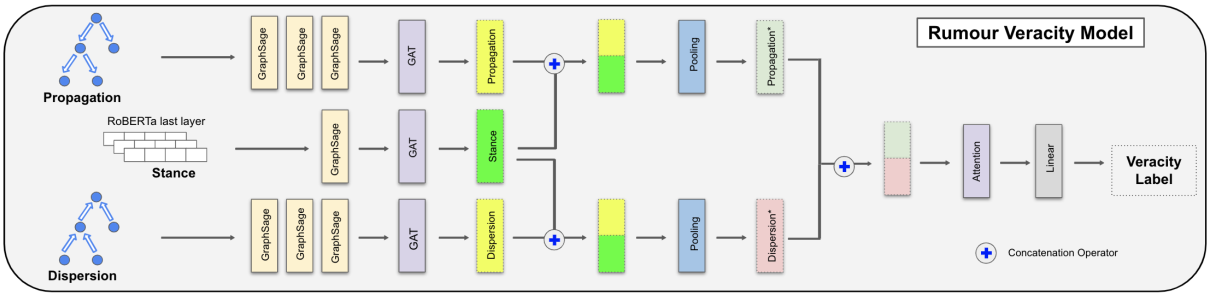

Figure 2: Architecture of our rumour verification model enhanced with structure-aware and stance-aware components based on graph neural networks. In the diagram, Propagation/Dispersion/Dispersion represent the outputs of each respective component, while Propagation*/Dispersion* represent the stance-enriched outputs of these.

	F	C	O	G	S	F1
Our model w/o stance	.228	.267	.300	.333	.293	.405
Our model with stance	.208	.341	.313	.403	.358	.434
SAVED (Dougrez-Lewis et al., 2021)	.372	.351	.304	.281	.332	.434

Table 1: PHEME results for each fold and overall reported as ~~macro~~ macro-averaged F1 ~~score~~scores. The fold abbreviations stand for Ferguson, Charlie Hebdo, Ottawa Shooting, Germanwings Crash and Sydney Siege

the best performing for rumour verification (Kochkina et al., 2018; Bian et al., 2020; Kochkina et al., 2023), given that the task heavily relies on user interactions for determining veracity. Our approach models the conversation thread as a graph, where interactions between posts manifest as propagation (top-down) and dispersion (bottom-up) flows similar to Bian et al. (2020). The architecture ~~is inspired by Bian et al. (2020). We replace the GCN with generalised~~ contains GraphSage (Hamilton et al., 2017) layers, proven to yield ~~more~~ meaningful node representations, followed by GAT (Veličković et al., 2018) layers, which are shown to improve performance in similar tasks (Kotonya et al., 2021; Zhang et al., 2021a; Jia et al., 2022). ~~The embeddings generated by the "all-MiniLM-L6-v2" model from Sentence Transformers~~ Sentence Transformers embeddings (Reimers and Gurevych, 2019) are used to initialise the node representations in the graphs. The propagation and dispersion component outputs are each concatenated with the output of a stance component and pooled, resulting in another concatenated representation to which a final multi-head attention layer (Vaswani et al., 2017) is applied.

Stance-Aware Component Stance detection is closely linked to misinformation detection (Hardalov et al., 2022) with previous work having shown that a joint approach improves rumour verification (Zubiaga et al., 2016; Derczynski et al., 2017; Gorrell et al., 2019; Yu et al., 2020; Dougrez-Lewis et al., 2021). As such our model includes a stance component unlike the GNN by Bian et al. (2020). Since only a small portion of the PHEME dataset is annotated with gold stance labels for the RumourEval (Task 8) competition (Derczynski et al., 2017), we generate silver labels for the whole corpus. In particular, we train a RoBERTa model (Liu et al., 2019) for stance classification and extract the embeddings from the last hidden layer to augment the rumour verification task with stance information. See Appendix D for experimental setup.

Performance of Rumour Verification Baselines We include the performance of our proposed baselines, the structure-aware model and its stance-aware version, in Table 1. ~~We report the macro-averaged F1-scores.~~

As expected, integrating stance knowledge into the model boosts performance by almost 3 F1-points overall with improved scores across the majority of folds; we hypothesise performance does not improve for the Ferguson fold due to its severe label imbalance skewed towards unverified rumours. Moreover, we observe that the model enhanced with the stance-aware component achieves competitive results and is comparable to the current state-of-the-art model on the PHEME dataset, the SAVED model by Dougrez-Lewis et al. (2021).

~~Experimental SetupWe train the rumour verification model for 300 epochs with learning rate 10^{-5}. The training loss is cross-entropy. The optimizer algorithm is Adam (Kingma and Ba, 2015). Hidden channel size is set as 256 for the propagation and~~

~~dispersion components and 32 hidden channel size for the stance component. The batch size is 20. For the GraphSage layers, we apply a mean aggreggator scheme, followed by a relu activation. For the Multi-headed Attention layer, we use 8 heads. To avoid overfitting, we randomly dropout an edge in the graph networks with probability 0.1. We use a Nvidia A5000 GPU for our model training. All model implementation is done via the *pytorch-geometric* package (Fey and Lenssen, 2019) for graph neural networks.~~

3.2 Explaining the Model

~~**Attribution Method**~~

3.2.1 Attribution Method

We experiment with two classes of attribution methods: gradient-based and game-theory-based. For gradient-based methods, we choose Integrated Gradients (IG) (Sundararajan et al., 2017). This is a local explainability algorithm that calculates attribution scores for each input ~~component (token for seq2seq models and node for graph neural networks)~~ unit by accumulating gradients along the interpolated path between a local point and a starting point with no information (baseline). ~~Integrated Gradients~~ IG was selected over other gradient-based saliency methods such as DeepLIFT (Shrikumar et al., 2017) as it has been shown to be more robust (Pruthi et al., 2022) when applied in classification tasks. Shapley Values (SV) (Štrumbelj and Kononenko, 2014) is the representative explainability method derived from game theory ~~. It has been previously~~ and has been used in many applications (Zhang et al., 2020; Mosca et al., 2021; Mamta and Ekbal, 2022). Its attribution scores ~~for each input feature~~ are calculated as expected marginal contributions where each feature is viewed as a 'player' within a ~~cooperative game-theory~~ coalitional game setting.

Note that we focus on post-hoc methods instead of intrinsic ones~~such as the attentionlayers~~, such as attention, in our architecture to keep the framework generalisable to other rumour verification models. Specifically, we use ~~Integrated Gradients and Shapley Values~~IG and SV[2] as methods for $\mathcal{E}$ to calculate the post importance f in Equation 1. This importance with respect to model prediction is then leveraged to sort the posts within the thread in descending order. We then construct subsets of important posts $\mathcal{I}_k \subset \mathcal{I}$ such that $|\mathcal{I}_k| = k\%|\mathcal{I}|$ with $\mathcal{I}_k$ representing the $k\%$ most important posts of the rumour thread, $k = 25, 50, 100$. These will be used as inputs for summarisation in the next stage to determine the trade-off between post importance and number of posts necessary to construct a viable explanation.

~~**Summarisation for Explanation** Abstractive summarisation has a dual purpose in~~

3.2.2 Summarisation for Explanation

We propose explanation baselines spanning different generation strategies: extractive vs abstractive, model-centric vs model-independent and in-domain vs out-of-domain.

Extractive Explanations

- *Important Response*: the response within the thread scored as most important by each attribution method. This is a model-dependent explanation.
- *Similar Response*: the response within the thread most semantically similar to the source claim, as scored by SBERT (Reimers and Gurevych, 2019). This model-independent baseline is inspired by (?).

Abstractive explanations have a dual purpose that fits the challenging set-up of our pipeline: ~~it serves~~ they serve as a way to aggregate important parts of the ~~conversation~~ thread, and ~~it also provides a fluent explanation to the model's prediction, as opposed to rationale-type explanations. We train a BART (Lewis et al., 2020) model~~ also provide a fluent justification sourced from multiple views to a claim's veracity.

- *Summary of $\mathcal{I}$*: We summarise the set $\mathcal{I}$ of important posts to obtain a model-centric explanation. We fine-tune BART (Lewis et al., 2020) on the MOS corpus introduced by Bilal et al. (2022) that addresses summarisation of topical groups of tweets by prioritising the majority opinion expressed. We hypothesise this template-guided[3] approach will satisfy the explanatory purpose since user opinion is an important indicator for assessing a

[2]~~The implementation of both methods is based on the~~ Used *captum* package (Kokhlikyan et al., 2020) for both.

[3]The template summary takes the form: *Main Story* + *Majority* Opinion expressed in the thread.

claim's veracity in rumour verification (Hardalov et al., 2022). ~~We generate summaries using as input the sets of posts $\mathcal{I}_{25}$, $\mathcal{I}_{50}$ and $\mathcal{I}$ to determine the trade-off between post importance and number of posts necessary to construct a viable explanation. We additionally consider an extractive explanation baseline that consists of the most important post within the set of responses to the source claim~~Similarly, we define explanations *Summary of $\mathcal{I}_{25}$* & *Summary of $\mathcal{I}_{50}$*.

- *Out-of-domain Summary*: We use the BART (Lewis et al., 2020) pre-trained on the CNN/Daily Mail (?) dataset without any fine-tuning and summarise the entire thread. This yields a model-independent explanation.

We note that while supervised summarisation is used to inform our generation strategy, our resulting explanations never rely on gold explanations annotated for the downstream task of fact-checking. In fact, neither MOS (Bilal et al., 2022) nor the CNN/Daily Mail (?) datasets were aimed for fact-checking and both focus on broad topics unrelated to the PHEME claims.

4 Automatic Evaluation of Explanation Quality

You will be shown a **Claim** and an **Explanation**. The veracity of the Claim can either be true, false or unverified. Choose an option from A to D that answers whether the Explanation can help confirm the veracity of the Claim.

A: The Explanation confirms the information in the Claim is true. The Explanation will include evidence to prove the Claim or show users believing the Claim.

B: The Explanation confirms the information in the Claim is false. The Explanation will include evidence to disprove the Claim or show users denying the Claim.

C: The Explanation confirms the information in the Claim is unverified. The Explanation will state no evidence exists to prove or disprove the Claim or show users doubting the Claim.

D: The Explanation is irrelevant in confirming the veracity of the Claim. The Explanation will not include any mention of evidence and users will not address the veracity of the Claim.

Claim: {claim}
Explanation: {explanation}

Table 2: Example task instructions used in the prompt ~~with fixed reasonings for each possible choice~~following a multiple-choice setting.

~~Evaluation is conducted to extrinsically assess the quality of the explanation summaries~~ As the PHEME dataset lacks gold standard explanations to compare against, we prioritise the extrinsic evaluation of the generated explanations with respect to their usefulness in downstream tasks. ~~We~~ This is similar to other work on explanatory fact-checking (Stammbach and Ash, 2020; Krishna et al., 2022).

In particular, we use the criterion of **informativeness** defined by Atanasova et al. (2020) as the ability to deduce the veracity of a claim based on the explanation. If the provided explanation is not indicative of any veracity label (*true*, *false*, or *unverified*), the explanation is considered uninformative. Otherwise, we compare the veracity suggested by the explanation to the prediction made by the model. This enables the evaluation of the explanation's fidelity to the model and ~~represents~~ is one of the main approaches to assess explanatory **faithfulness** in the research community (Jacovi and Goldberg, 2020).

We devise a novel evaluation strategy for capturing the informativeness of generated explanations based on ~~Large Language Models (LLMs)~~LLMs. This is motivated by recent work demonstrating the effectiveness of LLM reasoning capability in various tasks (Kojima et al., 2022; Chen, 2023), including as a zero-shot evaluator for summarisation outputs (Liu et al., 2023; Shen et al., 2023; Wang et al., 2023a). We use OpenAI's *gpt-3.5-turbo-0301*[4], hereinafter referred to as ChatGPT, which is a snapshot of the model from 1 March 2023 that will not receive updates – this should encourage the reproducibility of our evaluation. We follow ~~Shen et al. (2023) in providing fixed reasoning for each possible answer in the prompt , so as to prevent model hallucination~~a multiple-choice setting in the prompt similar to Shen et al. (2023). Our initial experiments confirmed previous findings (Brown et al., 2020) that GPT reasoning can be improved by including a few annotated ~~demonstration~~ representative examples of the evaluation within its prompt (~~i.e., in-context learning). Ten representative examples covering diverse explanation styles and veracity labels are selected (~~See Appendix A). We experimented with several prompt designs varying in level of detail (no justification of answer, no examples) and found that the most exhaustive prompt yielded best results. The final task instructions used for the prompt are ~~shown~~ in Table 2.

~~In line with Shen et al. (2023), we set the temperature parameter to 0 for reproducibility and send each request independently~~We ran a

[4] ~~We use~~ Used GPT-3.5-turbo due to its lower running costs compared to GPT-4.

pilot study (See Appendix C) to establish which temperature setting yields the most robust LLM evaluation. To account for any non-deterministic behaviour, the experiment was run three times. We find the results remain 100% consistent across runs for temperature 0. As this is in line with the settings used in similar works employing LLMs as evaluators (Shen et al., 2023), we also use this value for our experiment. Each request is sent independently via the Open AI API. Since using an LLM evaluator allows us to scale our evaluation (Chiang and Lee, 2023), we use all suitable PHEME threads[5] (i.e. ~~1,233 out of 2,107~~ 1233 / 2107 threads) for testing. This set-up foregoes the costs necessary to obtain a diverse manually-annotated test set and offers more statistical power to the results as recommended by Bowman and Dahl (2021).

5 Results and Discussion

~~Our evaluation is~~ The results are shown in Table 3. ~~The first row shows the % of uninformative explanations by explanation type. The rest show the % of explanations per type that yield (agree) or not (disagree) the same veracity prediction by ChatGPT as the rumour verification model. In each cell, the top scores are produced by Integrated Gradients and the bottom ones by Shapley Values.~~

Model-centric vs Model-independent We note that the explanations *Out-of-domain Summary* and *Similar Response* are independent of the rumour verification model built in section 3.1 as they are not produced by any of the post-hoc algorithms. Hence, while these are not expected to be faithful, we analyse how they compare in informativeness to the other model-centric explanations. We find that abstractive explanations (*Summaries of $\mathcal{I}_{25}$, $\mathcal{I}_{50}$, $\mathcal{I}$*) informed by the rumour verifier are the most informative of all. Thus, summarising a selection of important posts learned during the rumour verification process yields a better explanation than relying on individual replies or summarising the whole thread.

Integrated Gradients vs Shapley Values The summaries generated via ~~Integrated Gradients~~ IG achieve better scores than the ~~Shapley~~ SV ones in both informativeness and faithfulness. While

	~~Important Response~~ **Uninformative**	~~Summary of $\mathcal{I}_{25}$~~ **Unfaithful**	~~Summary of $\mathcal{I}_{50}$~~ **Faithful**
	~~Summary of $\mathcal{I}$~~	Extractive	
~~**Uninformative**~~ Important Response (IG)	67.23	**21.33**	11.44
Important Response (SV)	65.29	~~23.68~~ ~~22.95~~ 22.30	~~**22.11**~~ ~~23.60~~ 12.41
Similar Response	~~24.90~~ ~~23.60~~ 30.98	43.88	25.14
~~**Disagrees**~~	~~**21.33**~~ ~~22.30~~	Abstractive	
Summary of $\mathcal{I}_{25}$ (IG)	23.68	46.55	29.76
Summary of $\mathcal{I}_{25}$ (SV)	22.95	48.50	28.55
Summary of $\mathcal{I}_{50}$ (IG)	**22.11**	46.47 ~~47.20~~	~~48.58~~ ~~48.90~~ **30.41**
~~**Agrees**~~ Summary of $\mathcal{I}_{50}$ (SV)	~~11.44~~ ~~12.41~~ 23.60	~~29.76~~ ~~28.55~~ 47.20	~~**30.41**~~ 29.20
Summary of $\mathcal{I}$ (IG)	24.90	48.58	26.52
Summary of $\mathcal{I}$ (SV)	23.60	48.90	27.49
Out-of-domain Summary	39.17	38.28	22.55

Table 3: Explanation evaluation wrt model prediction (%). ~~Columns denote~~ If the explanation ~~type~~ cannot be used to infer a veracity label for the claim, it is **uninformative**. ~~'Agrees' means ChatGPT+~~ Otherwise, the explanation ~~matches~~ is **faithful** if its label coincides with the model ~~'s~~ prediction and ~~'Disagrees' is the opposite~~ **unfaithful** if not. Best scores are in bold.

~~Shapley~~ SV initially provides a better ~~Important Response~~ *Important Response*, it fails to detect other important posts within the thread as suggested by the scores for I_{25} and I_{50}. Moreover, the time complexity for the ~~Shapley~~ SV algorithm is exponential as its sampling strategy increases proportionately with the number of perturbed input permutations. We note the average computation time ~~(measured in seconds)~~ for both algorithms to assess a thread of 15 posts: 0.5s for ~~Integrated Gradients~~ IG and 2011s for ~~Shapley~~ SV. This makes ~~Integrated Gradients~~ IG a more suitable algorithm with respect to both performance ~~as well as~~ and running time.

Extractive Explanation The ~~extractive baseline explanation (post ranked as most important) is uninformative with respect to rumour veracity in two thirds of the cases (67.23% for IG and 65.29% for Shapley Values).~~ best extractive baseline is the *Similar Response*, which selects the closest semantic match from the thread to the claim. Followed by are model-centric baselines *Important Response* for both IG and SV, lagging behind by a large margin. We investigate the reason behind this performance by checking the

[5]Suitable defined as at least ten posts and the majority are non-empty after URL and user mentions are removed.

Claim
Update from Ottawa: Cdn soldier dies from shooting -Parliamentary guard wounded Parliament Hill still in lockdown URL

Prediction: Unverified

Explanation Summaries

Important Response: @TorontoStar Ok, time to take it to the *** muslims. Look out Allah, here comes the revenge. ***. (*Uninformative*)
Similar Response: #AttackinOttawa @TorontoStar: Update Cdn soldier dies from shooting -Parliamentary guard wounded Parliament Hill still in lockdown URL (*True*)

Summary of $\mathcal{I}_{25}$ (IG): Soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. The majority think the media is wrong to report that Parliament Hill was in lockdown and that the lockdown was a ploy to target Muslims. (*False*)

Summary of $\mathcal{I}_{50}$ (IG): Cdn soldier dies from shooting dead in Ottawa. The majority are sceptical about the news of the shooting and some are questioning where the confirmation is coming from. (*Unverified*)
Summary of $\mathcal{I}$ (IG): Cdn soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. Most users ask where the news of the gunman is and are wondering who is responsible for his death. Many of the responses use humour and irony, such as: 'I don't think the soldier is dead'. (*Unverified*)
Out-of-domain Summary: Update from Ottawa: -Cdn soldier dies from shooting -Parliamentary guard wounded. It looks like confirmations are coming in now. I don't think the soldier is dead. (*Unverified*)

Table 4: Example explanation summaries. ~~Red~~ Manually-annotated red highlights explain the model prediction for the given claim. ChatGPT evaluations are in ().

stance labels of the corresponding posts. Using the labelled data from Derczynski et al. (2017), we train a binary RoBERTa to identify comments and non-comments[6] where a comment is defined as a post that is unrelated or does not contribute to a rumour's veracity. ~~The original task is a 4-way classification of posts into one of the stance labels: *support*, *deny*, *query*, or *comment*. This is simplified by aggregating the first three labels into one.~~ We find that 64% of posts corresponding to ~~extractive explanations~~ *Important Response* labelled as uninformative are also classified as comments, much higher than 47% for *Similar Response*. This explains why ~~highest ranking posts alonecannot constitute suitable explanations (see also 'Most important response'in~~ semantic similarity can uncover a more relevant explanation than the *Important Response* alone. Still, this method suffers from 'echoing' the claim [7], which risks missing out on other important information found in the thread (see Table 4).

Abstract Explanation The abstractive explanations are shown to be considerably more informative than ~~the highest ranked responsemost~~ most extractive baselines. They have the advantage of aggregating useful information that appears later in the conversation. For instance, the ~~explanation summaries~~ abstractive explanations in Table 4 indicate ~~users~~posters' doubt and requests for more details. Furthermore, using an opinion-driven summariser is better for constructing a more informative summary-explanation than otherwise (See Sec. 3.2). We have also investigated the degree of information decay in relation to the number of posts used for summary construction ~~. We see in~~ in model-centric explanations. In Table 3~~that~~, the summary based on the first half of ~~the~~ important posts ($\mathcal{I}_{50}$) yields the most informative and faithful explanation for both algorithms, closely followed by the $\mathcal{I}_{25}$ ~~baseline~~one. The worst-performing ~~abstractive~~ model-centric explanation is that generated from the whole set of important replies ($\mathcal{I}$)[8]. We calculate the cumulative importance score of these data partitions and ~~observe that~~ note $\mathcal{I}_{25}$ and $\mathcal{I}_{50}$ contain ~~75% and 93%~~ 75% and 93% respectively of the thread's total importance. This suggests ~~that~~ the remaining second half of the importance-ordered thread offers little relevant information towards the model's decision.

6 Human Evaluation of LLM-based Evaluators

Agreement	Informativeness Detection	Veracity Prediction
~~A1~~ Ann - ~~A2~~ Ann	~~83~~82%	~~87~~88%
~~A1~~ Ann - ChatGPT	~~73~~69%	~~66~~68%
~~A2 - ChatGPT 63% 66%~~ ~~A1~~ Ann - ChatGPT 0613	~~67~~64%	~~65~~74%
~~A2 - ChatGPT 0613 71% 79%~~ ~~A1~~ Ann - GPT-4 ~~A2 - GPT-4 64% 82%~~	~~66~~63%	~~72~~80%

Table 5: Pairwise agreement scores for the overlap between the evaluations of ~~Annotator 1~~ the annotators (~~A1~~Ann) ~~, Annotator 2 (A2)~~ and the LLM. The LLMs are: ChatGPT ("gpt-3.5-turbo-0301"), ChatGPT 0613 ("gpt-3.5-turbo-0613") and GPT-4. The evaluations are conducted for two tasks: informativeness detection and veracity prediction.

Our human evaluation study has two goals: 1) quantify the evaluation capability of ChatGPT, the LLM employed in our ~~previous experiments in Section~~ experiments in Sec. 5 to assess automatic explanations and 2) investigate the performance of

[6]The original task is a 4-way classification of posts into one of the stance labels: *support*, *deny*, *query*, or *comment*. This is simplified by aggregating the first three labels into one.
[7]The majority of informative *Similar Responses* are classified as supporting the claim.

[8]~~For robustness of the results, we additionally calculate the agreement scores in relative terms for each explanation type, i.e. cases when the explanation matches the prediction out of the number of informative explanations. This complements Table 3 where the scores are reported with respect to the dataset size. We observe that the the rankings of the explanation types remains consistent in both relative and absolute settings.~~

ChatGPT against more recently-published LLMs. The results ~~of the human evaluations are found~~ are in Table 5.

We ran a pilot study on 50 threads randomly sampled, such that each fold and each label type is equally represented for a fair evaluation of the LLM performance. We follow a similar evaluation setup to the work of (Atanasova et al., 2020), who study whether their generated summaries provide support to the user in fact checking a claim. We ~~sense-check~~ check the LLM-based evaluation of automatic explanations on two tasks: 1. **Informativeness Detection**, where an Explanation is classified as either informative or uninformative and 2. **Veracity Prediction**, where an Informative Explanation is assigned true, false or unverified if it helps determine the veracity of the given claim.

Two Computer Science PhD candidates proficient in English were recruited as annotators for both tasks. Each annotator evaluated the ~~whole~~ test set of explanation candidates, resulting in ~~200~~ 300 evaluations per annotator. The same guidelines included in the prompt from Table 2 and manually-annotated examples from Appendix A are used as instructions to annotators. Before starting, the research team met with the annotators to ensure the tasks were understood, a process which lends itself to a richer engagement ~~between the human evaluators and the~~ with the guidelines.

6.1 Evaluation of ChatGPT

Informativeness Detection In our first human experiment (Table 5: first ~~row~~column), we evaluate whether ChatGPT correctly identifies an informative explanation. We find that the agreement between our annotators is ~~83~~82% ~~(i.e., 166 out of 200 summaries were given the same label by both annotators),~~ which we set as the upper threshold for comparison. We note that the agreement between human evaluators and ChatGPT consistently remains above the random baseline, but experiences a drop. Fleiss Kappa is ~~$\kappa = 0.447$~~$\kappa = 0.441$, which is higher than the agreement of $\kappa = 0.269, 0.345, 0.399$ reported by Atanasova et al. (2020) for the same binary setup. After examining the confusion matrix for this task (See Appendix B), it is observed that most mismatches arise from false positives ~~—~~ - ChatGPT labels an Explanation as informative when it is not. ~~Upon further inspection, we conclude that~~ Finally, we find this type of disagreement occurs in instances when the rumour is a complex claim, i.e., a claim with more than one check-worthy piece of information within it. As suggested by Chen et al. (2022), the analysis of complex real-world claims is a challenging task in the field of fact checking and we also observe its impact on our LLM-based evaluation for rumour verification.

Veracity Prediction In our second human experiment (Table 5: second ~~row~~column), we evaluate ~~whether~~ if ChatGPT correctly assigns a veracity label to an Informative explanation. Again, we consider ~~87~~88%, ~~which is~~ the agreement between our two annotators on this task, to be the upper threshold. Despite the more challenging set-up (ternary classification instead of binary), the LLM maintains good agreement: Fleiss Kappa ~~$\kappa = 0.434$~~ $\kappa = 0.451$ (again higher than those of Atanasova et al. (2020) for the multi-class setup $\kappa = 0.200, 0.230, 0.333$). Manual inspection of the disagreement cases reveals that the most frequent error type (~~38 out of 55~~ 58 / 75 mislabelled cases exhibit this pattern ~~—~~ - See Appendix B~~for more details~~) is when ChatGPT classifies a rumour as unverified based on the Explanation, while the annotator marks it as true. We hypothesise that an LLM fails to pick up on subtle cues present in the explanation that are otherwise helpful for deriving a veracity assessment. For instance, the Explanation *~~Ferguson police chief comes under scrutiny for his handling~~ "I think channel 7 news is saying he [the hostage-taker] is getting agitated bcoz of it [the ~~case~~hostage's escape], its time to go in.~~The majority believe that the chief is incompetent and his actions reveal his disregard for truth and justice~~"* ~~questions the motive of the police in the Ferguson case~~implies that the escape indeed took place as validated by Channel 7; this cue helps the annotator assign a true label to the corresponding claim *~~Anybody else thinks~~ "A sixth hostage has escaped from the ~~Ferguson police chief is just making this up as he goes along? This is beyond embarrassing. It's shameful~~Lindt cafe in Sydney!"*.

We ~~note that our ChatGPT-human agreement scores for both tasks are similar or higher to those reported by Zubiaga et al. (2016) who employ crowd-sourced workers for annotating similar classification subtasks on PHEME dataset: 61.1% for labelling certainty of rumours and 60.8% for classifying types of evidence arising from the thread. We~~ acknowledge the limitations of using an LLM as an evaluator, which reduces the richness of annotator interaction with the task, but show through our human evaluations that good agreement between an LLM and humans can still be achieved~~thus allowing .~~. This not only allows the scaling of final results to the entire dataset instead of being confined to a small test set (See ~~Section~~ 4~~).~~ Sec. 4), but also provides an automated benchmarking of generated explanations when the ground truth is missing.

6.2 Comparison to other LLMs

As ChatGPT is a closed-source tool continually updated by its team, it is important to investigate how ChatGPT-powered evaluations are influenced by the release of newer versions of the same language model or by substitution with improved models. To this effect, we compare the legacy version of ChatGPT released on 1 March 2023 with its more recent version, ChatGPT 0613 (released on 13 June 2023) and finally with GPT-4, a multimodal model equipped with broader general knowledge and more advanced reasoning capabilities.

We note that that while there are differences between the labels produced by the two versions, there is a higher agreement with human judgement for the newer snapshot ChatGPT 0613 ~~, especially~~ when assessed on the more ~~difficult~~ complex task of veracity prediction. A similar behaviour ~~can be~~ is observed for GPT-4, whose performance is the most aligned with human judgment in the second task. After examining the error patterns (See Appendix B), we observe a notable difference between ChatGPT-based models and GPT-4: while both temporal snapshots of ChatGPT tend to evaluate irrelevant explanations as informative (See ~~Subsection~~ Sec. 6.1), ~~the~~ GPT-4 ~~model~~ suffers from assigning too many false negatives. This implies the existence of a positive bias for ChatGPT models and a negative bias for GPT-4.

~~We hypothesise based~~ Based on our limited findings, we hypothesise that more recent models have the potential to be more reliable evaluators of ~~generated~~ explanations than older models, given their higher agreement with human annotators. However, the model choice needs to be grounded into the task requirements (i.e., which errors should be prioritised) and availability of computational costs (at the moment of writing GPT-4 is 20x more expensive than ChatGPT).

7 ~~Conclusion~~ Conclusions and Future Work

We ~~present a novel unsupervised~~ presented a novel zero-shot approach for generating abstractive explanations of model predictions for rumour verification. Our results ~~show that~~ showed abstractive summaries constructed from important posts scored by a post-hoc explainer algorithm can be successfully used to derive a veracity prediction given a claim and significantly outperform ~~an extractive baseline~~extractive and model-independent baselines. We also ~~find that~~ found using an LLM-based evaluator for assessing the quality of the generated summaries yields good agreement with human annotators for the tasks of informativeness detection and veracity prediction.

In future work, we plan to jointly train the veracity prediction and explanation generation ~~components jointly~~ and assess how an end-to-end ~~training~~ approach impacts the ~~faithfulness of generated explanationscompared to a modular approach as presented in this paper~~quality of resulting explanations. Additionally, we aim to enrich the explanations by incorporating external ~~heterogeneous~~ sources of information such as ~~news articles from~~ PHEMEPlus (Dougrez-Lewis et al., 2022). Another ~~future direction to pursue~~ direction is generating fine-grained explanations for addressing all check-worthy aspects within complex claims~~ – this requires the re-annotation of PHEME which is currently annotated at source claim level~~.

Limitations

Summarisation of threads The format of the conversation threads is challenging to summarise. Our approach to summarisation is to flatten the conversation tree and to concatenate the individual posts, which are then used as an input to a BART model. This approach is naïve as the meaning of the nested replies can be lost if considered independently of the context.

Task limitation At the moment, the explanations are constructed exclusively from the information present in the thread. Consequently, the degree of evidence present in a thread is reflected into the explanatory quality of the summary.

Complex Claims As seen in the paper, complex claims are a challenging subset of rumours to evaluate. Using the heuristic outlined in Chen et al. (2022) to identify complex claims based on verb count, we find that 22% of the claims within PHEME are classified as complex. To generate comprehensive explanations covering each check-worthy aspect within such claims, a re-annotation of PHEME is required which is only labelled at claim-level at the moment.

Human Evaluation Evaluation via large language models is in its infancy. While there have been very encouraging recent results of using it as a viable alternative to human evaluation, these are still early days. It is unclear how much the ~~model~~ evaluation stability is impacted by prompt design or by substitution with open-source language models.

~~Task limitation~~Evaluation criteria for generated output ~~At the moment,~~ Since our explanations rely on generation mechanisms including automatic summarisers, it is important to acknowledge that there are other evaluation criteria native to the generation field which are outside the scope of this paper and have not been covered. We note that since hallucination, redundancy, coherence and fluency have already been tested in the original works (Lewis et al., 2020; Bilal et al., 2022) introducing the summarisers we employ, we prioritised the criteria relevant to explainable fact-checking in the ~~explanations are constructed exclusively from the information present in the thread. Consequently, the degree of evidence present in a thread is reflected into the explanatory quality of the summary~~experiments of this paper: informativeness of explanations and faithfulness to predicted veracity label.

Ethics Statement

Our experiments use PHEME dataset, ~~which has already obtained~~ was given ethics approval upon its original release. However, we note that the dataset contains many instances of hate speech that may corrupt the intended aim of the summaries. In particular, summaries that use the majority of posts within the thread may exhibit hate-speech content exhibited by parts of the input text.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming. *Conference for Truth and Trust Online*.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):549–556.

Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. 2022. Template-based abstractive microblog opinion summarization. *Transactions of the Association for Computational Linguistics*, 10:1229–1248.

Alodie Boissonnet, Marzieh Saeidi, Vassilis Plachouras, and Andreas Vlachos. 2022. Explainable assessment of healthcare articles with QA. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

John Dougrez-Lewis, Elena Kochkina, Miguel Arana-Catania, Maria Liakata, and Yulan He. 2022. PHEMEPlus: Enriching social media rumour verification with external evidence. In *Proceedings of the Fifth Fact Extraction and VERification Workshop (FEVER)*, pages 49–58, Dublin, Ireland. Association for Computational Linguistics.

John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. 2021. Learning disentangled latent topics for Twitter rumour veracity classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3902–3908, Online. Association for Computational Linguistics.

Matthias Fey and Jan E. Lenssen. 2019. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation?

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Hao Jia, Honglei Wang, and Xiaoping Zhang. 2022. Early detection of rumors based on source tweet-word graph attention networks. *PLoS One*, 17(7):e0271224.

Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing. *Information*, 13(10).

Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and abstractive explanations for fact-checking and evaluation of news. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 45–50, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Elena Kochkina, Tamanna Hossain, Robert L. Logan, Miguel Arana-Catania, Rob Procter, Arkaitz Zubiaga, Sameer Singh, Yulan He, and Maria Liakata. 2023. Evaluating the generalisability of neural rumour verification models. *Information Processing and Management*, 60(1):103116.

Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Neema Kotonya, Thomas Spooner, Daniele Magazzeni, and Francesca Toni. 2021. Graph reasoning with context-aware linearization for interpretable fact extraction and verification. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 21–30, Dominican Republic. Association for Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

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. *CoRR*, abs/1907.11692.

. Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 478–492, Online only. Association for Computational Linguistics.

Edoardo Mosca, Maximilian Wich, and Georg Groh. 2021. Understanding and interpreting the impact of user context in hate speech detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 91–102, Online. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Rob Procter, Miguel Arana-Catania, Yulan He, Maria Liakata, Arkaitz Zubiaga, Elena Kochkina, and Runcong Zhao. 2023. Some observations on fact-checking work with implications for computational support.

Rob Procter, Farida Vis, and Alex Voss. 2013. Reading the riots on twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*, 16(3):197–214.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating Explanations: How Much Do Explanations from the Teacher Aid Students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization?

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries forautomated fact checking. In *Conference for Truth and Trust Online*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations*.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023. Less is more for long document summary evaluation by llms.

Jianfei Yu, Jing Jiang, Ling Min Serena Khoo, Hai Leong Chieu, and Rui Xia. 2020. Coupled hierarchical transformer for stance-aware rumor verification in social media conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1392–1401, Online. Association for Computational Linguistics.

Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang. 2020. An explainable machine learning framework for fake financial news detection. In *Proceedings of the 41st International Conference on Information Systems, ICIS 2020, Making Digital Inclusive: Blending the Locak and the Global, Hyderabad, India, December 13-16, 2020*. Association for Information Systems.

Xinpeng Zhang, Shuzhi Gong, and Richard O. Sinnott. 2021a. Social media rumour detection through graph attention networks. In *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, pages 1–6.

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021b. Faxplainac: A fact-checking tool based on explainable models with human correction in the loop. CIKM '21, page 4823–4827, New York, NY, USA. Association for Computing Machinery.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLOS ONE*, 11:1–29.

A Examples of Assessing the Informativeness of Explanations

B Error Analysis of LLM's performance as Evaluator

We note that our ChatGPT-human agreement scores for both tasks are similar or higher to those reported by Zubiaga et al. (2016), who employ crowd-sourced workers for annotating similar classification subtasks on the PHEME dataset: 61.1% for labelling certainty of rumours and 60.8% for classifying types of evidence arising from the thread.

We report the performance of ChatGPT, ChatGPT 0614 and GPT-4 as evaluators using the manually annotated set of 200 explanations. The error analysis is shared via a confusion matrix for each task: informativeness detection (See Table 7) and veracity prediction (See Table 8). The results are reported as counts.

C Pilot Study on Temperature Setting for ChatGPT

We used the same explanations in Table 4 and ran a small pilot study to assess how incrementing the temperature parameter affects the LLM evaluation. Results are in Table 9. We used increments of 0.2 in temperature and ran the experiment 3 times to account for the non-deterministic behaviour. Overall, the evaluations remain consistent (94% of the labels output by ChatGPT are the same) across runs and temperature values. In particular, we note that when using temperature 0, the evaluations remain 100% consistent and for non-zero temperature, the evaluation only impacts the labelling of the last explanation which is less helpful than previous explanation candidates.

D Experimental Setup

We train the rumour verification model for 300 epochs with learning rate 10^{-5}. The training loss is cross-entropy. The optimizer algorithm is Adam (Kingma and Ba, 2015). Hidden channel size is set as 256 for the propagation and dispersion components and 32 hidden channel size for the stance component. The batch size is 20. For the GraphSage layers, we apply a mean aggreggator scheme, followed by a relu activation. For the Multi-headed Attention layer, we use 8 heads. Embeddings generated by the "all-MiniLM-L6-v2" model from Sentence Transformers (Reimers and Gurevych, 2019) are used to initialise the node representations in the graphs. To avoid overfitting, we randomly

Claim: Victims were forced to hold a flag on the cafe window.
Explanation: Users believe this is true and point to the released footage.
Your answer: A

Claim: BREAKING: Hostages are running out of the cafe #sydneysiege
Explanation: Some users believe the claim is unverified as Channel 9 did not confirm and some agree that the details of potential escape should not be disclosed.
Your answer: C

Claim: One of the gunmen left an ID behind in the car.
Explanation: One of the gunmen left an ID behind in the car. The majority deny the ID was found there and point to the media for blame.
Your answer: B

Claim: Three people have died in the shooting.
Explanation: Three people have died in the shooting. Most users pray the attack is over soon.
Your answer: D

Claim: NEWS #Germanwings co-pilot Andreas Lubitz had serious depressive episode (Bild newspaper) #4U9525 URL LINK
Explanation:Germanwings co-pilot Andrés Lubitz has serious depressive episode. Never trust bild. Users believe that bild is a fake newspaper and the stories concerned with the suicide of Andreas Lubitz should not be discussed.
Your answer: C

Claim: Snipers set up on National Art Gallery as we remain barricaded in Centre Block on Parliament Hill #cdnpoli.
Explanation: Snipers set up on National Art Gallery as we remain barricaded in Centre Block on Parliament Hill. Most users are skeptical about the news and await more details.
Your answer: C

Claim: BREAKING: #Germanwings co-pilot's name is Andreas Lubitz, a German national, says Marseilles prosecutor.
Explanation: He didn't have a political or religious background.
Your answer: D

Claim: Several bombs have been placed in the city
Explanation: This is false, why then cause panic and circulate on social media?
Your answer: B

Claim: Police report the threats released by the criminals.
Explanation: The majority threaten to condemn anyone who is a terrorist.
Your answer: D

Claim: #CharlieHebdo attackers shouted 'The Prophet is avenged'.
Explanation: In video showing assassination of officer.walking back to car they shouted: 'we avenged the prophet.We killed Charlie Hebdo'
Your answer: A

Table 6: ~~Examples of assessing informativeness of explanations~~Ten representative examples covering diverse explanation styles and veracity labels are selected. These are included in the final prompt for ChatGPT.

dropout an edge in the graph networks with probability 0.1. We use a Nvidia A5000 GPU for our model training. All model implementation is done via the *pytorch-geometric* package (Fey and Lenssen, 2019) for graph neural networks.

Annotator / LLM	Informative	Uninformative
	ChatGPT	
Informative	~~163~~ 169	~~85~~ 107
Uninformative	~~45~~ 81	~~107~~ 143
	ChatGPT 0613	
Informative	~~155~~ 236	~~83~~ 104
Uninformative	~~53~~ 114	~~109~~ 146
	GPT-4	
Informative	~~91~~ 160	~~25~~ 30
Uninformative	~~117~~ 190	~~167~~ 220

Table 7: Confusion Matrices for ChatGPT, ChatGPT 0613 and ChatGPT-4 for the task of **Informativeness Detection**

Annotator / LLM	True	False	Unverified
		ChatGPT	
True	~~50~~ 105	~~2~~ 3	4
False	~~7~~ 12	~~13~~ 18	~~2~~ 5
Unverified	~~38~~ 58	~~2~~ 3	~~45~~ 61
		ChatGPT 0613	
True	~~71~~ 114	~~2~~ 3	8
False	~~7~~ 10	10	~~4~~ 6
Unverified	~~14~~ 26	~~2~~ 8	~~35~~ 51
		GPT-4	
True	~~38~~ 78	0	~~1~~ 2
False	~~6~~ 10	~~8~~ 10	9
Unverified	~~5~~ 7	~~0~~ 84	~~24~~ 40

Table 8: Confusion Matrices for ChatGPT, ChatGPT 0613 and ChatGPT-4 for the task of **Veracity Prediction**

Explanation	$T = 0$	$T = 0.2$	$T = 0.4$	$T = 0.6$	$T = 0.8$	$T = 1$
@TorontoStar Ok, time to take it to the ***muslims. Look out Allah, here comes the revenge. ***.	D,D,D	D,D,D	D,D,D	D,D,D	D,D,D	D,D,D
Soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. The majority think the media is wrong to report that Parliament Hill was in lockdown and that the lockdown was a ploy to target Muslims.	B,B,B	B,B,B	B,B,B	B,B,B	B,B,B	B,B,B
Cdn soldier dies from shooting dead in Ottawa. The majority are sceptical about the news of the shooting and some are questioning where the confirmation is coming from.	C,C,C	C,C,C	C,C,C	C,C,C	C,C,C	C,C,C
Cdn soldier dies from shooting in Ottawa and Parliament Hill is in lockdown. Most users ask where the news of the gunman is and are wondering who is responsible for his death. Many of the responses use humour and irony, such as: 'I don't think the soldier is dead'.	C,C,C	C,A,C	C,C,C	C,C,C	C,A,A	C,C,A

Table 9: Labels output by ChatGPT for each explanations across 3 different runs.