

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

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

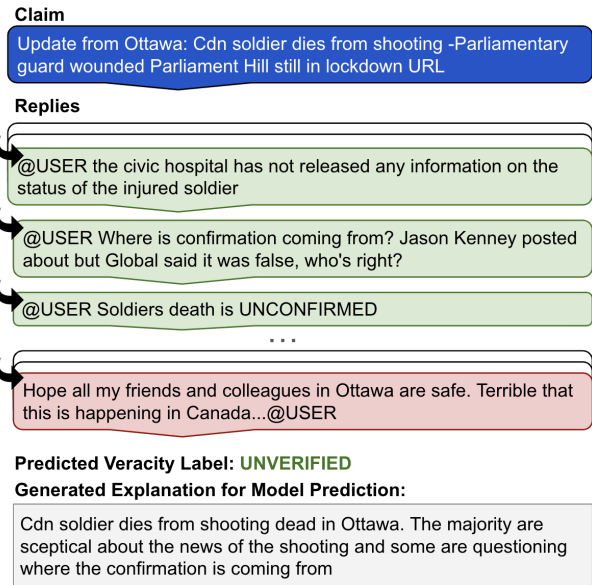


Figure 1: Example of a PHEME thread for which the claim is predicted to be *unverified* by our model. Our proposed pipeline first identifies replies which **agree** or **disagree** with the model prediction and then summarises the former ones to generate an explanation for the model’s prediction.

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 (see Figure 1). 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

057 generate model-centric explanation summaries of
058 the model’s assessments.

059 [Atanasova et al. \(2020\)](#), [Kotonya and Toni](#)
060 [\(2020\)](#) and [Stammbach and Ash \(2020\)](#) were the
061 first to introduce explanation summaries for fact-
062 checking across different datasets. [Kotonya and](#)
063 [Toni \(2020\)](#) provided a framework for creating
064 abstractive summaries that justify the true verac-
065 ity of the claim in the PUBHealth dataset, simi-
066 larly to [Stammbach and Ash \(2020\)](#) who augment
067 FEVER ([Thorne et al., 2018](#)) dataset with a corpus
068 of explanations. [Atanasova et al. \(2020\)](#) proposed a
069 jointly trained system that produces veracity predic-
070 tions and explanations extracted from ruling com-
071 ments on LIAR-PLUS ([Alhindi et al., 2018](#)). The
072 approach in ([Kotonya and Toni, 2020](#)) results in ex-
073 planatory summaries that are, however, not faithful
074 to the model, while [Atanasova et al. \(2020\)](#) requires
075 a supervised approach. Our goal is to create a novel
076 zero-shot method for abstractive explanations that
077 explain the rumour verification model’s predictions.
078 We make the following contributions:

- 079 • We introduce a zero-shot framework for gen-
080 erating abstractive explanations using opinion-
081 guided summarisation for the task of rumour ver-
082 ification. To the best of our knowledge, this is the
083 first time free-text explanations are introduced
084 for this task.
- 085 • We investigate the benefits of using a gradient-
086 based algorithm and a game theoretical algorithm
087 to provide explainability.
- 088 • While our explanation generation method is gen-
089 eralisable to any verification model, we introduce
090 a novel graph-based hierarchical approach.
- 091 • We evaluate the informativeness of several ex-
092 planation baselines, by providing them as input
093 to a few-shot trained large language model. Our
094 results show that our proposed abstractive model-
095 centric explanations are on average more infor-
096 mative in 77% of the cases as opposed to 49%
097 for all other baselines.
- 098 • We provide both human and LLM-based evalua-
099 tion of the generated explanations, showing that
100 LLMs achieve sufficient agreement with humans,
101 thus allowing scaling of the evaluation of the
102 explanatory summaries in absence of gold-truth
103 explanations.

104 2 Related Work

105 **Explainable Fact Checking** Following the ex-
106 ample of fact-checking organisations (e.g., Snopes,

107 Full Fact, Politifact), which provide journalist-
108 written justifications to determine the truthfulness
109 of claims, recent datasets augmented with free-
110 text explanations have been constructed: LIAR-
111 PLUS ([Alhindi et al., 2018](#)), PubHealth ([Kotonya](#)
112 [and Toni, 2020](#)), AVeriTeC ([Schlichtkrull et al.,](#)
113 [2023](#)). A wide range of explainable outputs and
114 methods have been proposed: theorem proofs ([Kr-](#)
115 [ishna et al., 2022](#)), knowledge graphs ([Ah-](#)
116 [madi et al., 2019](#)), question-answer decomposi-
117 tions ([Boissonnet et al., 2022](#); [Chen et al., 2022](#)),
118 reasoning programs ([Pan et al., 2023](#)), deployable
119 evidence-based tools ([Zhang et al., 2021b](#)) and sum-
120 marisation ([Atanasova et al., 2020](#); [Kotonya et al.,](#)
121 [2021](#); [Stammbach and Ash, 2020](#); [Kazemi et al.,](#)
122 [2021](#); [Jolly et al., 2022](#)). We adopt summarisation
123 as our generation strategy as it fluently aggregates
124 evidence from multiple inputs and has been proven
125 effective in similar works which we discuss next.

Explainability as Summarisation [Atanasova](#)
126 [et al. \(2020\)](#) and [Kotonya and Toni \(2020\)](#) lever-
127 aged large-scale datasets annotated with gold jus-
128 tifications to generate supervised explanations for
129 fact-checking, while [Stammbach and Ash \(2020\)](#)
130 used few-shot learning on GPT-3 to create the e-
131 FEVER dataset of explanations. Similar to ([Stamm-](#)
132 [bach and Ash, 2020](#)), [Kazemi et al. \(2021\)](#) also
133 leveraged a GPT-based model (GPT-2) to gener-
134 ate abstractive explanations, but found that that
135 their extractive baseline, Biased TextRank, out-
136 performed GPT-2 on the LIAR-PLUS dataset ([Al-](#)
137 [hindi et al., 2018](#)). [Jolly et al. \(2022\)](#) warn that
138 the output of extractive explainers lacks fluency
139 and sentential coherence, which motivated their
140 work on unsupervised post-editing using the ex-
141 planations produced by [Atanasova et al. \(2020\)](#).
142 Our approach is different from the above as we
143 derive our summaries from microblog content (as
144 opposed to news articles as done by [Atanasova et al.](#)
145 [\(2020\)](#); [Stammbach and Ash \(2020\)](#); [Kazemi et al.](#)
146 [\(2021\)](#); [Jolly et al. \(2022\)](#)), and only use the subset
147 of posts relevant to the model’s decision to inform
148 the summary (rather than summarising the whole
149 input as in ([Kotonya and Toni, 2020](#); [Kazemi et al.,](#)
150 [2021](#)). Moreover, we rely on a zero-shot generation
151 approach without gold explanations, contrary to
152 ([Atanasova et al., 2020](#); [Kotonya and Toni, 2020](#)).153

LLMs as evaluators Having generated explana-
154 tory summaries the question arises as to how to
155 evaluate them at scale. LLMs have been employed
156 as knowledge bases for fact-checking ([Lee et al.,](#)
157

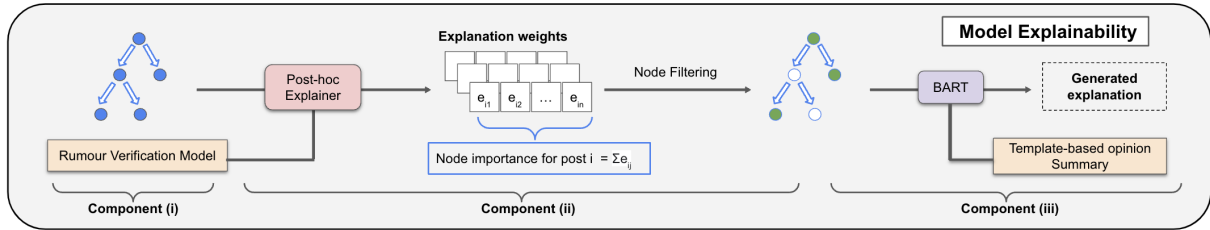


Figure 2: Framework of our proposed approach to obtain faithful generated explanations for the rumour verifier.

2020; Pan et al., 2023), as explanation generators
 for assessing a claim’s veracity (Stammach and
 Ash, 2020; Kazemi et al., 2021) and, as of recently,
 as evaluators in generation tasks. Most works fo-
 cused 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 us-
 ing LLM-based evaluators, our evaluation is most
 similar to Liu et al. (2023a); Shen et al. (2023); Chi-
 ang 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 faith-
 fulness of explanations for verifying a claim.

3 Methodology

Our methodological approach (Figure 2) consists
 of three individual components: *i*) training a ru-
 mour verification model; *ii*) using a post-hoc ex-
 plainability algorithm; *iii*) generating summary-
 explanations. The approach to explanation genera-
 tion is zero-shot and model-agnostic.

We demonstrate our approach on PHEME (Zubi-
 aga 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 training.

Task Formulation For a model trained on ru-
 mour verification \mathcal{M} , an attribution-based expla-
 nation method \mathcal{E} , and a rumourous conversation
 thread consisting of posts $\mathcal{T} = \{p_1, \dots, p_l\}$ with
 embeddings $\{x_1, \dots, x_l\} \subset \mathbb{R}^n$, 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 $e_i \in \mathbb{R}^n$ is the attribution vector for embed-
 ding x_i of post p_i such that each value e_{ij} cor-
 responds 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 pre-
 diction, 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 pre-
 diction (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 3), which caters for a flexible information
 structure combining information in the conversa-
 tion 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 mod-
 els such as tree-based and graph-based are among
 the best performing for rumour verification (Kochk-
 ina et al., 2018; Bian et al., 2020; Kochkina et al.,
 2023), given that the task heavily relies on user in-
 teractions 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 sim-
 ilar to Bian et al. (2020). The architecture con-
 tains 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

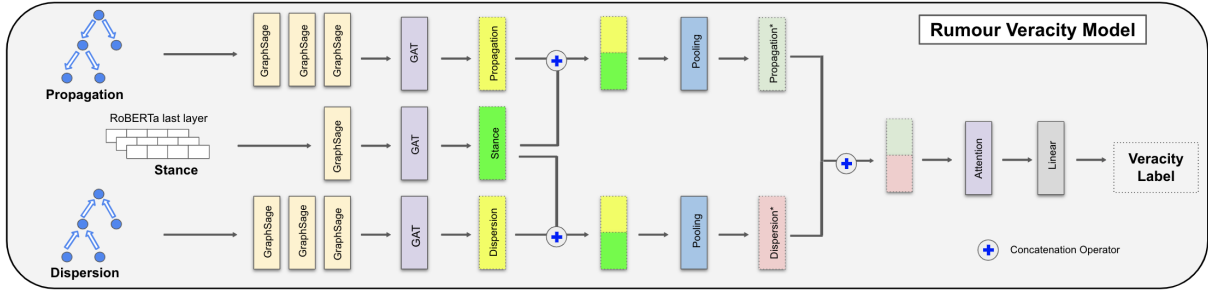


Figure 3: Architecture of our GNN-based rumour verifier enhanced with structure and stance-aware components: Propagation/Dispersion represent the outputs of each respective component, while Propagation*/Dispersion* represent the stance-enriched outputs of these.

(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 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.

| | F | C | O | G | S | F1 |
|------------------------------------|------|------|------|------|------|------|
| Model w/o stance & dispersion | .235 | .241 | .281 | .372 | .371 | .360 |
| Model w/o stance | .228 | .267 | .300 | .333 | .293 | .405 |
| Model with stance & dispersion | .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

Ablation study for Rumour Verifier We include a short ablation study of our proposed baselines in

Table 1. As expected, removing both stance and structure knowledge from the model degrades performance by almost 7 F1-points overall. The model enhanced with all the components (stance, propagation and dispersion) outperforms its counterparts 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, our complete model 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 architec-

298 ture to keep the framework generalisable to other
 299 rumour verification models. Specifically, we use
 300 IG and SV² as methods for \mathcal{E} to calculate the post
 301 importance f in Eq 1. This importance with respect
 302 to model prediction is then leveraged to sort the
 303 posts within the thread in descending order. We
 304 then construct subsets of important posts $\mathcal{I}_k \subset \mathcal{I}$
 305 such that $|\mathcal{I}_k| = k\%|\mathcal{I}|$ with \mathcal{I}_k representing the
 306 $k\%$ most important posts of the rumour thread,
 307 $k = 25, 50, 100$. These will be used as inputs for
 308 summarisation in the next stage to determine the
 309 trade-off between post importance and number of
 310 posts necessary to construct a viable explanation.

3.2.2 Summarisation for Explanation

312 We propose explanation baselines spanning dif-
 313 ferent generation strategies: extractive vs abstrac-
 314 tive, model-centric vs model-independent and in-
 315 domain vs out-of-domain.

Extractive Explanations

- 317 • *Important Response*: the response within the
 318 thread scored as most important by each attri-
 319 bution method. This is model-dependent.
- 320 • *Similar Response*: the response within the thread
 321 most semantically similar to the source claim, as
 322 scored by SBERT (Reimers and Gurevych, 2019).
 323 This model-independent baseline is inspired by
 324 (Russo et al., 2023).

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

- 330 • *Summary of \mathcal{I}* : We summarise the set \mathcal{I} of impor-
 331 tant posts to obtain a model-centric explanation.
 332 We fine-tune BART (Lewis et al., 2020) on the
 333 MOS corpus by Bilal et al. (2022) that addresses
 334 summarisation of topical groups of tweets by pri-
 335 oritising the majority opinion expressed. Both
 336 MOS and PHEME were collected from the same
 337 platform, Twitter. We hypothesise this template-
 338 guided³ approach will satisfy the explanatory pur-
 339 pose since user opinion is an important indicator
 340 for assessing a claim’s veracity in rumour ver-
 341 ification (Hardalov et al., 2022). Similarly, we
 342 define explanations *Summary of $\mathcal{I}_{25}, \mathcal{I}_{50}$* .
 343 • *Out-of-domain Summary*: We use the BART
 344 (Lewis et al., 2020) pre-trained on the CNN/

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

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

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 result-
 ing 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

356 As the PHEME dataset lacks gold standard explana-
 357 tions to compare against, we prioritise the extrinsic
 358 evaluation of the generated explanations with re-
 359 spect to their usefulness in downstream tasks. This
 360 is similar to work on explanatory fact-checking
 361 (Stammach and Ash, 2020; Krishna et al., 2022).
 362

363 In particular, we use the criterion of **informa-**
 364 **tiveness** defined by Atanasova et al. (2020) as the
 365 ability to deduce the veracity of a claim based on
 366 the explanation. If the provided explanation is not
 367 indicative of any veracity label, the explanation is
 368 considered uninformative. Otherwise, we compare
 369 the veracity suggested by the explanation to the
 370 prediction made by the model. This enables the
 371 evaluation of the explanation’s fidelity to the model
 372 and is one of the main approaches to assess explana-
 373 tory **faithfulness** (Jacovi and Goldberg, 2020).
 374

375 We devise a novel evaluation strategy for cap-
 376 turing the informativeness of generated explana-
 377 tions based on LLMs. This is motivated by recent
 378 work demonstrating the effectiveness of LLMs as
 379 zero-shot reasoners and judges for various tasks
 380 (Kojima et al., 2022; Chen, 2023; Chan et al., 2023;
 381 Zheng et al., 2023), including as a zero-shot evalu-
 382 ator for summarisation outputs (Liu et al., 2023b;
 383 Shen et al., 2023; Wang et al., 2023a; Liu et al.,
 384 2023a). We use OpenAI’s *gpt-3.5-turbo-0301*⁴,
 385 hereinafter referred to as ChatGPT. We follow a
 386 multiple-choice setting in the prompt similar to
 387 Shen et al. (2023). Our initial experiments con-
 388 firmed previous findings (Brown et al., 2020) that
 389 GPT reasoning can be improved by including a few
 390 annotated representative examples of the evaluation

⁴Used GPT-3.5-turbo due to its lower running costs com-
 pared to GPT-4. This is a snapshot of the model from 1 March
 2023 that will not receive updates – this should encourage the
 reproducibility of our evaluation.

within its prompt ⁵ (See Appx. A).

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 threads⁶ (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 2.

| | 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 2: 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}* ,

⁵We 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

⁶Suitable 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: @USER Ok, time to take it to the *** muslims. Look out Allah, here comes the revenge. ***. (Uninformative) |
| Similar Response: #AttackinOttawa @USER: 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 3: Example explanation summaries. Manually-annotated red highlights explain the model prediction for the given claim. ChatGPT evaluations are in ().

\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 \mathcal{I}_{25} and \mathcal{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-

comments⁷ 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 claim⁸, which risks missing out on other important information found in the thread (see Table 3).

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 3 indicate posters’ doubt and requests for more details. Furthermore, using an opinion-driven summariser is better for constructing a more informative 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 2, 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

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 4.

We ran a pilot study on 50 threads randomly sampled, such that each fold and each label type

⁷The 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.

⁸The majority of informative *Similar Responses* are classified as supporting the claim.

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

Table 4: 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.

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 and examples from Appendix A are used as instructions.

6.1 Evaluation of ChatGPT

Informativeness Detection In our first human experiment (Table 4: 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 moderate but 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 4: 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 multi-modal 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.

634 Limitations

635 **Summarisation of threads** The format of the
636 conversation threads is challenging to summarise.
637 Our approach to summarisation is to flatten the
638 conversation tree and to concatenate the individual
639 posts, which are then used as an input to a BART
640 model. This approach is naïve as the meaning of
641 the nested replies can be lost if considered indepen-
642 dently of the context. We posit that a graph-based
643 neural summariser capable of encoding both the
644 hierarchy of posts and their information as nodes
645 in a graph, would benefit the summarisation of mi-
646 croblog opinions. A relevant approach has been
647 proposed by Wang et al. (2020) for extractive sum-
648 marisation of news articles, though this would need
649 to be adapted for the more challenging format of
650 microblog posts and account for the potential topic
651 shifts exhibited by very long threads as seen in Sun
652 and Loparo (2019).

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

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

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

675 **Evaluation criteria for generated output** Since
676 our explanations rely on generation mechanisms
677 including automatic summarisers, it is important to
678 acknowledge that there are other evaluation crite-
679 ria native to the generation field which are outside
680 the scope of this paper and have not been covered.
681 We note that since hallucination, redundancy, co-
682 herence 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: infor-
mativeness of explanations and faithfulness to pre-
dicted 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 Mo-
hammed Saeed. 2019. Explainable fact checking
with probabilistic answer set programming. *Confer-
ence for Truth and Trust Online*.
- Tariq Alhindi, Savvas Petridis, and Smaranda Mure-
san. 2018. **Where is your evidence: Improving fact-
checking by justification modeling**. In *Proceedings
of the First Workshop on Fact Extraction and VERi-
fication (FEVER)*, pages 85–90, Brussels, Belgium.
Association for Computational Linguistics.
- Pepa Atanasova, Jakob Grue Simonsen, Christina Li-
oma, 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 sum-
marization**. *Transactions of the Association for Com-
putational 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 Compu-
tational Linguistics.
- Samuel R. Bowman and George Dahl. 2021. **What will
it take to fix benchmarking in natural language under-
standing?** In *Proceedings of the 2021 Conference of*

| | | | |
|-----|-------------------------------------------------------------------------------|------------------------------------------------------------------------------|-----|
| 849 | Elena Kochkina, Tamanna Hossain, Robert L. Logan, | Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, | 905 |
| 850 | Miguel Arana-Catania, Rob Procter, Arkaitz Zubi- | Ruo Chen Xu, and Chenguang Zhu. 2023a. G-eval: | 906 |
| 851 | aga, Sameer Singh, Yulan He, and Maria Liakata. | Nlg evaluation using gpt-4 with better human align- | 907 |
| 852 | 2023. Evaluating the generalisability of neural ru- | ment. | 908 |
| 853 | mour verification models. <i>Information Processing</i> | | |
| 854 | and Management , 60(1):103116. | | |
| 855 | Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. | Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- | 909 |
| 856 | 2018. All-in-one: Multi-task learning for rumour | Joshi, Danqi Chen, Omer Levy, Mike Lewis, | 910 |
| 857 | verification. In <i>Proceedings of the 27th International</i> | Luke Zettlemoyer, and Veselin Stoyanov. 2019. | 911 |
| 858 | <i>Conference on Computational Linguistics</i> , pages | Roberta: A robustly optimized BERT pretraining | 912 |
| 859 | 3402–3413, Santa Fe, New Mexico, USA. Associa- | approach. <i>CoRR</i> , abs/1907.11692. | 913 |
| 860 | tion for Computational Linguistics. | | |
| 861 | Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu- | Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan | 914 |
| 862 | taka Matsuo, and Yusuke Iwasawa. 2022. Large lan- | Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, | 915 |
| 863 | guage models are zero-shot reasoners. In <i>Advances</i> | Feng Sun, and Qi Zhang. 2023b. Calibrating llm- | 916 |
| 864 | in Neural Information Processing Systems. | based evaluator. | 917 |
| 865 | Narine Kokhlikyan, Vivek Miglani, Miguel Martin, | . Mamta and Asif Ekbal. 2022. Adversarial sample | 918 |
| 866 | Edward Wang, Bilal Alsallakh, Jonathan Reynolds, | generation for aspect based sentiment classification. | 919 |
| 867 | Alexander Melnikov, Natalia Kliushkina, Carlos | In <i>Findings of the Association for Computational</i> | 920 |
| 868 | Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. | <i>Linguistics: ACL-IJCNLP 2022</i> , pages 478–492, | 921 |
| 869 | Captum: A unified and generic model interpretability | Online only. Association for Computational Linguis- | 922 |
| 870 | library for pytorch. | tics. | 923 |
| 871 | Neema Kotonya, Thomas Spooner, Daniele Magazzeni, | Edoardo Mosca, Maximilian Wich, and Georg Groh. | 924 |
| 872 | and Francesca Toni. 2021. Graph reasoning with | 2021. Understanding and interpreting the impact of | 925 |
| 873 | context-aware linearization for interpretable fact ex- | user context in hate speech detection. In <i>Proceedings</i> | 926 |
| 874 | traction and verification. In <i>Proceedings of the</i> | of the Ninth International Workshop on Natural Lan- | 927 |
| 875 | <i>Fourth Workshop on Fact Extraction and VERification</i> | guage Processing for Social Media , pages 91–102, | 928 |
| 876 | <i>(FEVER)</i> , pages 21–30, Dominican Republic. | Online. Association for Computational Linguistics. | 929 |
| 877 | Association for Computational Linguistics. | | |
| 878 | Neema Kotonya and Francesca Toni. 2020. Explainable | Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, | 930 |
| 879 | automated fact-checking for public health claims. In | Caglar Gulcehre, and Bing Xiang. 2016. Abstrac- | 931 |
| 880 | <i>Proceedings of the 2020 Conference on Empirical</i> | tive text summarization using sequence-to-sequence | 932 |
| 881 | <i>Methods in Natural Language Processing (EMNLP)</i> , | RNNs and beyond. In <i>Proceedings of the 20th</i> | 933 |
| 882 | pages 7740–7754, Online. Association for Computa- | <i>SIGNLL Conference on Computational Natural Lan-</i> | 934 |
| 883 | tional Linguistics. | <i>guage Learning</i> , pages 280–290, Berlin, Germany. | 935 |
| 884 | Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. | Association for Computational Linguistics. | 936 |
| 885 | 2022. ProFVer: Natural logic theorem proving for | Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan | 937 |
| 886 | fact verification. <i>Transactions of the Association for</i> | Luu, William Yang Wang, Min-Yen Kan, and Preslav | 938 |
| 887 | <i>Computational Linguistics</i> , 10:1013–1030. | Nakov. 2023. Fact-checking complex claims with | 939 |
| 888 | Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau | program-guided reasoning. In <i>Proceedings of the</i> | 940 |
| 889 | Yih, Hao Ma, and Madian Khabsa. 2020. Language | <i>61st Annual Meeting of the Association for Compu-</i> | 941 |
| 890 | models as fact checkers? In <i>Proceedings of the</i> | <i>tational Linguistics (Volume 1: Long Papers)</i> , pages | 942 |
| 891 | <i>Third Workshop on Fact Extraction and VERification</i> | 6981–7004, Toronto, Canada. Association for Com- | 943 |
| 892 | <i>(FEVER)</i> , pages 36–41, Online. Association for | putational Linguistics. | 944 |
| 893 | Computational Linguistics. | | |
| 894 | Mike Lewis, Yinhan Liu, Naman Goyal, Marjan | Rob Procter, Miguel Arana-Catania, Yulan He, Maria | 945 |
| 895 | Ghazvininejad, Abdelrahman Mohamed, Omer Levy, | Liakata, Arkaitz Zubiaga, Elena Kochkina, and | 946 |
| 896 | Veselin Stoyanov, and Luke Zettlemoyer. 2020. | Runcong Zhao. 2023. Some observations on fact- | 947 |
| 897 | BART: Denoising sequence-to-sequence pre-training | checking work with implications for computational | 948 |
| 898 | for natural language generation, translation, and com- | support. | 949 |
| 899 | prehension. In <i>Proceedings of the 58th Annual Meet-</i> | Rob Procter, Farida Vis, and Alex Voss. 2013. Reading | 950 |
| 900 | <i>ing of the Association for Computational Linguistics</i> , | the riots on twitter: methodological innovation for | 951 |
| 901 | pages 7871–7880, Online. Association for Computa- | the analysis of big data. <i>International Journal of</i> | 952 |
| 902 | tional Linguistics. | <i>Social Research Methodology</i> , 16(3):197–214. | 953 |
| 903 | Zachary C. Lipton. 2018. The mythos of model inter- | Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, | 954 |
| 904 | pretability. <i>Commun. ACM</i> , 61(10):36–43. | Livio Baldini Soares, Michael Collins, Zachary C. | 955 |
| | | Lipton, Graham Neubig, and William W. Cohen. | 956 |
| | | 2022. Evaluating Explanations: How Much Do Ex- | 957 |
| | | planations from the Teacher Aid Students? <i>Transac-</i> | 958 |
| | | <i>tions of the Association for Computational Linguis-</i> | 959 |
| | | <i>tics</i> , 10:359–375. | 960 |

| | | |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| 961 | Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks . In <i>Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing</i> . Association for Computational Linguistics. | 1013 |
| 962 | | 1014 |
| 963 | | 1015 |
| 964 | | 1016 |
| 965 | | 1017 |
| 966 | Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023. Benchmarking the generation of fact checking explanations . <i>Transactions of the Association for Computational Linguistics</i> , 11:1250–1264. | 1018 |
| 967 | | 1019 |
| 970 | Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web . | 1020 |
| 971 | | 1021 |
| 972 | | 1022 |
| 973 | Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? | 1023 |
| 974 | | 1024 |
| 975 | | 1025 |
| 976 | Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In <i>International conference on machine learning</i> , pages 3145–3153. PMLR. | 1026 |
| 977 | | 1027 |
| 978 | | 1028 |
| 979 | | 1029 |
| 980 | | 1030 |
| 981 | Dominik Stammach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. In <i>Conference for Truth and Trust Online</i> . | 1031 |
| 982 | | 1032 |
| 983 | | 1033 |
| 984 | Yingcheng Sun and Kenneth Loparo. 2019. Topic shift detection in online discussions using structural context . In <i>2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)</i> , volume 1, pages 948–949. | 1034 |
| 985 | | 1035 |
| 986 | | 1036 |
| 987 | | 1037 |
| 988 | | 1038 |
| 989 | Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In <i>Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17</i> , page 3319–3328. JMLR.org. | 1039 |
| 990 | | 1040 |
| 991 | | 1041 |
| 992 | | 1042 |
| 993 | | 1043 |
| 994 | James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task . In <i>Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)</i> , pages 1–9, Brussels, Belgium. Association for Computational Linguistics. | 1044 |
| 995 | | 1045 |
| 996 | | 1046 |
| 997 | | 1047 |
| 998 | | 1048 |
| 999 | | 1049 |
| 1000 | | 1050 |
| 1001 | Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need . In <i>Advances in Neural Information Processing Systems</i> , volume 30. Curran Associates, Inc. | 1051 |
| 1002 | | 1052 |
| 1003 | | 1053 |
| 1004 | | 1054 |
| 1005 | | 1055 |
| 1006 | Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. <i>International Conference on Learning Representations</i> . | 1056 |
| 1007 | | 1057 |
| 1008 | | 1058 |
| 1009 | | 1059 |
| 1010 | Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions . <i>Knowl. Inf. Syst.</i> , 41(3):647–665. | 1060 |
| 1011 | | 1061 |
| 1012 | | 1062 |
| | Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization . In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 6209–6219, Online. Association for Computational Linguistics. | 1063 |
| | | 1064 |
| | | 1065 |
| | | 1066 |
| | 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 . | 1067 |
| | | 1068 |
| | | 1069 |
| | 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 . | 1070 |
| | | 1071 |
| | | 1072 |
| | Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023. Less is more for long document summary evaluation by llms . | 1073 |
| | | 1074 |
| | | 1075 |
| | 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 <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 1392–1401, Online. Association for Computational Linguistics. | 1076 |
| | | 1077 |
| | | 1078 |
| | | 1079 |
| | | 1080 |
| | Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang. 2020. An explainable machine learning framework for fake financial news detection. In <i>Proceedings of the 41st International Conference on Information Systems, ICIS 2020, Making Digital Inclusive: Blending the Local and the Global, Hyderabad, India, December 13-16, 2020</i> . Association for Information Systems. | 1081 |
| | | 1082 |
| | | 1083 |
| | | 1084 |
| | | 1085 |
| | | 1086 |
| | | 1087 |
| | | 1088 |
| | | 1089 |
| | | 1090 |
| | | 1091 |
| | | 1092 |
| | | 1093 |
| | | 1094 |
| | | 1095 |
| | | 1096 |
| | | 1097 |
| | | 1098 |
| | | 1099 |
| | | 1100 |
| | | 1101 |
| | | 1102 |
| | | 1103 |
| | | 1104 |
| | | 1105 |
| | | 1106 |
| | | 1107 |
| | | 1108 |
| | | 1109 |
| | | 1110 |
| | | 1111 |
| | | 1112 |
| | | 1113 |
| | | 1114 |
| | | 1115 |
| | | 1116 |
| | | 1117 |
| | | 1118 |
| | | 1119 |
| | | 1120 |
| | | 1121 |
| | | 1122 |
| | | 1123 |
| | | 1124 |
| | | 1125 |
| | | 1126 |
| | | 1127 |
| | | 1128 |
| | | 1129 |
| | | 1130 |
| | | 1131 |
| | | 1132 |
| | | 1133 |
| | | 1134 |
| | | 1135 |
| | | 1136 |
| | | 1137 |
| | | 1138 |
| | | 1139 |
| | | 1140 |
| | | 1141 |
| | | 1142 |
| | | 1143 |
| | | 1144 |
| | | 1145 |
| | | 1146 |
| | | 1147 |
| | | 1148 |
| | | 1149 |
| | | 1150 |
| | | 1151 |
| | | 1152 |
| | | 1153 |
| | | 1154 |
| | | 1155 |
| | | 1156 |
| | | 1157 |
| | | 1158 |
| | | 1159 |
| | | 1160 |
| | | 1161 |
| | | 1162 |
| | | 1163 |
| | | 1164 |
| | | 1165 |
| | | 1166 |
| | | 1167 |
| | | 1168 |
| | | 1169 |
| | | 1170 |
| | | 1171 |
| | | 1172 |
| | | 1173 |
| | | 1174 |
| | | 1175 |
| | | 1176 |
| | | 1177 |
| | | 1178 |
| | | 1179 |
| | | 1180 |
| | | 1181 |
| | | 1182 |
| | | 1183 |
| | | 1184 |
| | | 1185 |
| | | 1186 |
| | | 1187 |
| | | 1188 |
| | | 1189 |
| | | 1190 |
| | | 1191 |
| | | 1192 |
| | | 1193 |
| | | 1194 |
| | | 1195 |
| | | 1196 |
| | | 1197 |
| | | 1198 |
| | | 1199 |
| | | 1200 |

A Guidelines and Examples for Assessing the Informativeness of Explanations

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

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 300 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 LLM 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 aggregator 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

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 #sydneyseige
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.

| | | | | |
|------|--------------------------------------------------------|----------|-------------------------------------------------------------------------------------------------------------------------------|------|
| 1121 | training. All model implementation is done via | E | Current Submission colour-coded for the changes we have implemented compared to the previous version of the manuscript | 1124 |
| 1122 | the <i>pytorch-geometric</i> package (Fey and Lenssen, | | | 1125 |
| 1123 | 2019) for graph neural networks. | | | 1126 |
| | | | | 1127 |
| | | | Red stands for removed material and blue stands for new additions. | 1128 |
| | | | | 1129 |

| 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**

| 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.

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

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

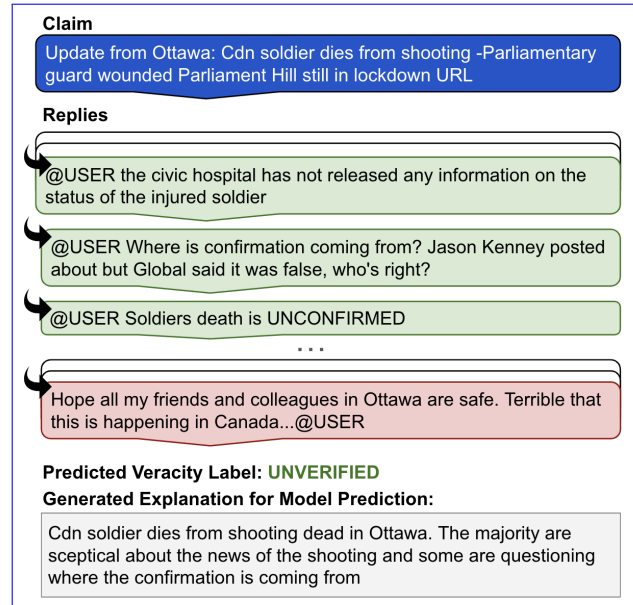


Figure 1: Example of a PHEME thread for which the claim is predicted to be *unverified* by our model. Our proposed pipeline first identifies replies which agree or disagree with the model prediction and then summarises the former ones to generate an explanation for the model’s prediction.

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 (see Figure 1). 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 em-

056 ploy the conversation threads that form its input to
057 generate model-centric explanation summaries of
058 the model’s assessments.

059 [Atanasova et al. \(2020\)](#), [Kotonya and Toni](#)
060 [\(2020\)](#) and [Stammbach and Ash \(2020\)](#) were the
061 first to introduce explanation summaries for fact-
062 checking across different datasets. [Kotonya and](#)
063 [Toni \(2020\)](#) provided a framework for creating ab-
064 stractive summaries that justify the true veracity
065 of the claim in the PUBHealth dataset, similarly
066 to [Stammbach and Ash \(2020\)](#) who augment the
067 FEVER ([Thorne et al., 2018](#)) dataset with a cor-
068 pus of explanations. [Atanasova et al. \(2020\)](#) pro-
069 posed a jointly trained system that produces verac-
070 ity predictions ~~as well as explanations in the form~~
071 ~~of extracted evidence and explanations extracted~~
072 from ruling comments on the LIAR-PLUS dataset
073 ([Alhindi et al., 2018](#)). The approach in ([Kotonya](#)
074 [and Toni, 2020](#)) results in explanatory summaries
075 that are, however, not faithful to the model, while
076 [Atanasova et al. \(2020\)](#) requires a supervised ap-
077 proach. Our goal is to create a novel zero-shot
078 method for abstractive explanations that explain
079 the rumour verification model’s predictions. We
080 make the following contributions:

- 081 • We introduce a zero-shot framework for gen-
082 erating abstractive explanations using opinion-
083 guided summarisation for the task of rumour ver-
084 ification. To the best of our knowledge, this is the
085 first time free-text explanations are introduced
086 for this task.
- 087 • We investigate the benefits of using a gradient-
088 based algorithm and a game theoretical algorithm
089 to provide explainability.
- 090 • While our explanation generation method is gen-
091 eralisable to any verification model, we introduce
092 a novel graph-based hierarchical approach.
- 093 • We evaluate the informativeness of several expla-
094 nation baselines, ~~including model-independent~~
095 ~~and model-dependent ones stemming from the~~
096 ~~highest scoring posts~~ by providing them as in-
097 put to a few-shot trained large language model.
098 Our results show that our proposed abstractive
099 model-centric explanations are on average more
100 informative in 77% of the cases as opposed to
101 49% for all other baselines.
- 102 • We provide both human and LLM-based evalua-
103 tion of the generated explanations, showing that
104 LLMs achieve sufficient agreement with humans,
105 thus allowing scaling of the evaluation of the
106 explanatory summaries in absence of gold-truth
107 explanations.

2 Related Work 108

Explainable Fact Checking Following the ex- 109
110 ample of fact-checking organisations (e.g., Snopes,
111 Full Fact, Politifact), which provide journalist-
112 written justifications to determine the truthfulness
113 of claims, recent datasets augmented with free-
114 text explanations have been constructed: LIAR-
115 PLUS ([Alhindi et al., 2018](#)), PubHealth ([Kotonya](#)
116 [and Toni, 2020](#)), AVeriTeC ([Schlichtkrull et al.,](#)
117 [2023](#)). A wide range of explainable outputs and
118 methods have been proposed: theorem proofs ([Kr-](#)
119 [ishna et al., 2022](#)), knowledge graphs ([Ah-](#)
120 [madi et al., 2019](#)), question-answer decomposi-
121 tions ([Boissonnet et al., 2022](#); [Chen et al., 2022](#)),
122 reasoning programs ([Pan et al., 2023](#)), deployable
123 evidence-based tools ([Zhang et al., 2021b](#)) and sum-
124 marisation ([Atanasova et al., 2020](#); [Kotonya et al.,](#)
125 [2021](#); [Stammbach and Ash, 2020](#); [Kazemi et al.,](#)
126 [2021](#); [Jolly et al., 2022](#)). We adopt summarisation
127 as our generation strategy as it fluently aggregates
128 evidence from multiple inputs and has been proven
129 effective in similar works which we discuss next.

Explainability as Summarisation [Atanasova](#) 130
131 [et al. \(2020\)](#) and [Kotonya and Toni \(2020\)](#) lever-
132 aged large-scale datasets annotated with gold jus-
133 tifications to generate supervised explanations for
134 fact-checking, while [Stammbach and Ash \(2020\)](#)
135 used few-shot learning on GPT-3 to create the e-
136 FEVER dataset of explanations. Similar to ([Stamm-](#)
137 [bach and Ash, 2020](#)), [Kazemi et al. \(2021\)](#) also
138 leveraged a GPT-based model (GPT-2) to gener-
139 ate abstractive explanations, but found that that
140 their extractive baseline, Biased TextRank, out-
141 performed GPT-2 on the LIAR-PLUS dataset ([Al-](#)
142 [hindi et al., 2018](#)). [Jolly et al. \(2022\)](#) warn that
143 the output of extractive explainers lacks fluency
144 and sentential coherence, which motivated their
145 work on unsupervised post-editing using the ex-
146 planations produced by [Atanasova et al. \(2020\)](#).
147 Our approach is different from the above as we
148 derive our summaries from microblog content (as
149 opposed to news articles as done by [Atanasova et al.](#)
150 [\(2020\)](#); [Stammbach and Ash \(2020\)](#); [Kazemi et al.](#)
151 [\(2021\)](#); [Jolly et al. \(2022\)](#)), and only use the subset
152 of posts relevant to the model’s decision to inform
153 the summary (rather than summarising the whole
154 input as in ([Kotonya and Toni, 2020](#); [Kazemi et al.,](#)
155 [2021](#)). Moreover, we rely on a zero-shot generation
156 approach without gold explanations, contrary to
157 ([Atanasova et al., 2020](#); [Kotonya and Toni, 2020](#)).

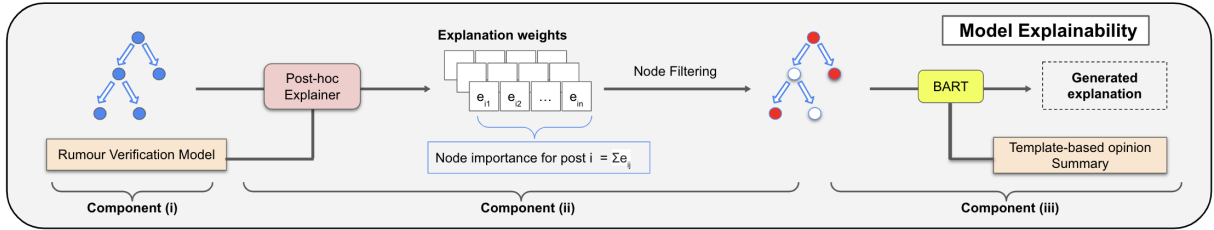


Figure 2: Framework of our proposed approach to obtain faithful generated explanations for the rumour verification model verifier. It explains the process of explanation generation, where the weights from a model are passed through an explainer algorithm to identify important input nodes, which are then filtered and used in abstractive summarisation.

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 (Stammach 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. (2023a); 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 2) consists of three individual components: *i*) training a rumour 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 training.

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 $\mathcal{T} = \{p_1, \dots, p_l\}$ with embeddings $\{x_1, \dots, x_l\} \subset \mathbb{R}^n$, 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 $e_i \in \mathbb{R}^n$ 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 3), 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.

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

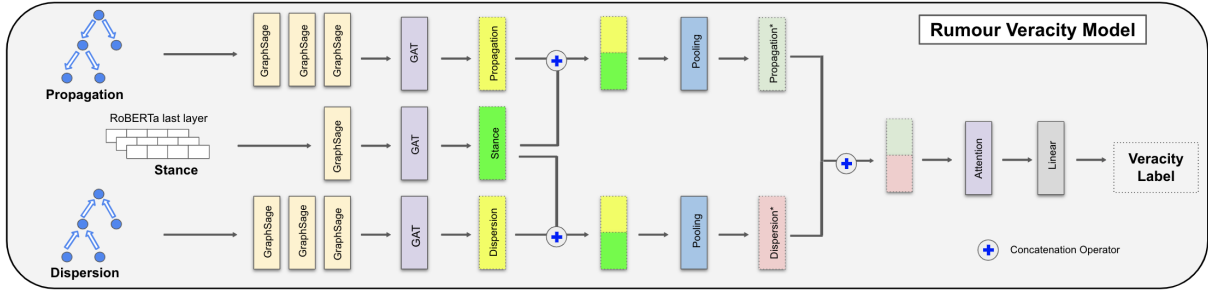


Figure 3: Architecture of our [GNN-based rumour verification model-verifier](#) enhanced with [structure-aware structure](#) and stance-aware components [based on graph neural networks](#). In the diagram, [/Propagation/Dispersion](#) represent the outputs of each respective component, while [Propagation*/Dispersion*](#) represent the stance-enriched outputs of these.

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.

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 Ablation study for Rumour Verification Baselines We include the performance-Verifier We include a short

| | F | C | O | G | S | F1 |
|------------------------------------------------|------|------|------|------|------|------|
| Model w/o stance & dispersion | .235 | .241 | .281 | .372 | .371 | .360 |
| Model w/o stance | .228 | .267 | .300 | .333 | .293 | .405 |
| Model with stance & dispersion | .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](#)

[ablation study](#) 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 removing both stance and structure knowledge from the model degrades~~ performance by almost 3-7 F1-points overall ~~with improved scores~~. ~~The model enhanced with all the components (stance, propagation and dispersion) outperforms its counterparts~~ 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 our complete model~~ 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 SV² as methods for \mathcal{E} to calculate the post importance f in Equation Eq 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.

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).

²Used *captum* package (Kohlikiyan et al., 2020) for both.

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. Both MOS and PHEME were collected from the same platform, Twitter. We hypothesise this template-guided³ 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 $\mathcal{I}_{50}, \mathcal{I}_{100}$* .

- *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 (Stammach 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.

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

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). (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 **LLMs as zero shot reasoners and judges for various tasks** (Kojima et al., 2022; Chen, 2023; Chan et al., 2023; Zheng et al., 2023), including as a zero-shot evaluator for summarisation outputs (Liu et al., 2023b; Shen et al., 2023; Wang et al., 2023a; Liu 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 Appendix⁵ (See Appx. 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 5.~~

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 threads⁶ (i.e. 1233 / 2107 threads)

⁴Used GPT-3.5-turbo due to its lower running costs compared to GPT-4. This is a snapshot of the model from 1 March 2023 that will not receive updates – this should encourage the reproducibility of our evaluation.

⁵We 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

⁶Suitable defined as at least ten posts and the majority are non-empty after URL and user mentions are removed.

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 2.

| | 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 2: 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 \mathcal{I}_{25} and \mathcal{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

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 3: [Example explanation summaries. Manually-annotated red highlights explain the model prediction for the given claim. ChatGPT evaluations are in \(\).](#)

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-comments⁷ 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 claim⁸, which risks missing out on other important information found in the thread (see Table 3).

Abstractive Explanation The abstractive explanations are shown to be considerably more informative than most extractive baselines. They

⁷The 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.

⁸The majority of informative *Similar Responses* are classified as supporting the claim.

have the advantage of aggregating useful information that appears later in the conversation. For instance, the abstractive explanations in Table 3 indicate posters’ doubt and requests for more details. Furthermore, using an opinion-driven summariser is better for constructing a more informative 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 2, 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 4: 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 4.

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

528 explanations on two tasks: 1. **Informativeness**
529 **Detection**, where an Explanation is classified as either
530 informative or uninformative and 2. **Veracity**
531 **Prediction**, where an Informative Explanation is
532 assigned true, false or unverified if it helps deter-
533 mine the veracity of the given claim.

534 Two Computer Science PhD candidates profi-
535 cient in English were recruited as annotators for
536 both tasks. Each annotator evaluated the test set of
537 explanation candidates, resulting in 300 evaluations
538 per annotator. The same guidelines ~~included in the~~
539 ~~prompt from Table 5~~ and examples from Appendix
540 ~~A~~ are used as instructions. ~~Before starting, the~~
541 ~~research team met with the annotators to ensure~~
542 ~~the tasks were understood, a process which lends~~
543 ~~itself to a richer engagement with the guidelines.~~

544 6.1 Evaluation of ChatGPT

545 **Informativeness Detection** In our first human
546 experiment (Table 4: first column), we evaluate
547 whether ChatGPT correctly identifies an informa-
548 tive explanation. We find that the agreement be-
549 tween our annotators is 82% which we set as the
550 upper threshold for comparison. We note that the
551 agreement between human evaluators and Chat-
552 GPT consistently remains above the random base-
553 line, but experiences a drop. Fleiss Kappa is
554 $\kappa = 0.441$, which is moderate but higher than
555 the agreement of $\kappa = 0.269, 0.345, 0.399$ reported
556 by Atanasova et al. (2020) for the same binary
557 setup. After examining the confusion matrix for
558 this task (See Appendix B), it is observed that most
559 mismatches arise from false positives - ChatGPT
560 labels an Explanation as informative when it is not.
561 Finally, we find this type of disagreement occurs in
562 instances when the rumour is a complex claim, i.e.,
563 a claim with more than one check-worthy piece of
564 information within it. As suggested by Chen et al.
565 (2022), the analysis of complex real-world claims
566 is a challenging task in the field of fact checking
567 and we also observe its impact on our LLM-based
568 evaluation for rumour verification.

569 **Veracity Prediction** In our second human exper-
570 iment (Table 4: second column), we evaluate if
571 ChatGPT correctly assigns a veracity label to an
572 Informative explanation. Again, we consider 88%,
573 the task annotator agreement to be the upper thresh-
574 old. Despite the more challenging set-up (ternary
575 classification instead of binary), the LLM main-
576 tains good agreement: Fleiss Kappa $\kappa = 0.451$
577 (again higher than those of Atanasova et al. (2020)

578 for the multi-class setup $\kappa = 0.200, 0.230, 0.333$).
579 Manual inspection of the disagreement cases re-
580 veals that the most frequent error type (58 / 75 mis-
581 labelled cases exhibit this pattern - See Appendix
582 B) is when ChatGPT classifies a rumour as unveri-
583 fied based on the Explanation, while the annotator
584 marks it as true. We hypothesise that an LLM fails
585 to pick up on subtle cues present in the explanation
586 that are otherwise helpful for deriving a veracity
587 assessment. For instance, the Explanation "*I think*
588 *channel 7 news is saying he [the hostage-taker] is*
589 *getting agitated bcoz of it [the hostage's escape],*
590 *its time to go in.*" implies that the escape indeed
591 took place as validated by Channel 7; this cue helps
592 the annotator assign a true label to the correspond-
593 ing claim "*A sixth hostage has escaped from the*
594 *Lindt cafe in Sydney!*".

595 We acknowledge the limitations of using an
596 LLM as an evaluator, which reduces the richness
597 of annotator interaction with the task, but show
598 through our human evaluations that good agree-
599 ment between an LLM and humans can still be
600 achieved. This not only allows the scaling of fi-
601 nal results to the entire dataset instead of being
602 confined to a small test set (See Sec. 4), but also
603 provides an automated benchmarking of generated
604 explanations when the ground truth is missing.

605 6.2 Comparison to other LLMs

606 As ChatGPT is a closed-source tool continually
607 updated by its team, it is important to investigate
608 how ChatGPT-powered evaluations are influenced
609 by the release of newer versions of the same lan-
610 guage model or by substitution with improved mod-
611 els. To this effect, we compare the legacy version
612 of ChatGPT released on 1 March 2023 with its
613 more recent version, ChatGPT 0613 (released on
614 13 June 2023) and finally with GPT-4, a multi-
615 modal model equipped with broader general knowl-
616 edge and more advanced reasoning capabilities.

617 We note that that while there are differences
618 between the labels produced by the two versions,
619 there is a higher agreement with human judgement
620 for the newer snapshot ChatGPT 0613 when as-
621 sessed on the more complex task of veracity pre-
622 diction. A similar behaviour is observed for GPT-4,
623 whose performance is the most aligned with hu-
624 man judgment in the second task. After examining
625 the error patterns (See Appendix B), we observe a
626 notable difference between ChatGPT-based mod-
627 els and GPT-4: while both temporal snapshots of
628 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. [We posit that a graph-based neural summariser capable of encoding both the hierarchy of posts and their information as nodes in a graph, would benefit the summarisation of microblog opinions. A relevant approach has been proposed by Wang et al. \(2020\) for extractive sum-](#)

[marisation of news articles, though this would need to be adapted for the more challenging format of microblog posts and account for the potential topic shifts exhibited by very long threads as seen in Sun and Loparo \(2019\).](#)

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

| | | |
|-----|----------------------------------------------------------------------|-----|
| 726 | hate-speech content exhibited by parts of the input | 781 |
| 727 | text. | 782 |
| 728 | References | |
| 729 | Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mo- | |
| 730 | hammed Saeed. 2019. Explainable fact checking | |
| 731 | with probabilistic answer set programming. <i>Confer-</i> | |
| 732 | <i>ence for Truth and Trust Online</i> . | |
| 733 | Tariq Alhindi, Savvas Petridis, and Smaranda Mure- | |
| 734 | san. 2018. Where is your evidence: Improving fact- | 783 |
| 735 | checking by justification modeling . In <i>Proceedings</i> | 784 |
| 736 | <i>of the First Workshop on Fact Extraction and VERi-</i> | 785 |
| 737 | <i>fication (FEVER)</i> , pages 85–90, Brussels, Belgium. | 786 |
| 738 | Association for Computational Linguistics. | |
| 739 | Pepa Atanasova, Jakob Grue Simonsen, Christina Li- | |
| 740 | oma, and Isabelle Augenstein. 2020. Generating fact | |
| 741 | checking explanations . In <i>Proceedings of the 58th</i> | |
| 742 | <i>Annual Meeting of the Association for Computational</i> | |
| 743 | <i>Linguistics</i> , pages 7352–7364, Online. Association | |
| 744 | for Computational Linguistics. | |
| 745 | Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing | |
| 746 | Huang, Yu Rong, and Junzhou Huang. 2020. Rumour | |
| 747 | detection on social media with bi-directional graph | |
| 748 | convolutional networks . <i>Proceedings of the AAAI</i> | |
| 749 | <i>Conference on Artificial Intelligence</i> , 34(01):549– | |
| 750 | 556. | |
| 751 | Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong | |
| 752 | Nguyen, Rob Procter, and Maria Liakata. 2022. | |
| 753 | Template-based abstractive microblog opinion sum- | |
| 754 | marization . <i>Transactions of the Association for Com-</i> | |
| 755 | <i>putational Linguistics</i> , 10:1229–1248. | |
| 756 | Alodie Boissonnet, Marzieh Saeidi, Vassilis Plachouras, | |
| 757 | and Andreas Vlachos. 2022. Explainable assessment | |
| 758 | of healthcare articles with QA . In <i>Proceedings of the</i> | |
| 759 | <i>21st Workshop on Biomedical Language Processing</i> , | |
| 760 | pages 1–9, Dublin, Ireland. Association for Compu- | |
| 761 | tational Linguistics. | |
| 762 | Samuel R. Bowman and George Dahl. 2021. What will | |
| 763 | it take to fix benchmarking in natural language under- | |
| 764 | standing? In <i>Proceedings of the 2021 Conference of</i> | |
| 765 | <i>the North American Chapter of the Association for</i> | |
| 766 | <i>Computational Linguistics: Human Language Tech-</i> | |
| 767 | <i>nologies</i> , pages 4843–4855, Online. Association for | |
| 768 | Computational Linguistics. | |
| 769 | Tom Brown, Benjamin Mann, Nick Ryder, Melanie | |
| 770 | Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind | |
| 771 | Neelakantan, Pranav Shyam, Girish Sastry, Amanda | |
| 772 | Askell, Sandhini Agarwal, Ariel Herbert-Voss, | |
| 773 | Gretchen Krueger, Tom Henighan, Rewon Child, | |
| 774 | Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens | |
| 775 | Winter, Chris Hesse, Mark Chen, Eric Sigler, Ma- | |
| 776 | teusz Litwin, Scott Gray, Benjamin Chess, Jack | |
| 777 | Clark, Christopher Berner, Sam McCandlish, Alec | |
| 778 | Radford, Ilya Sutskever, and Dario Amodei. 2020. | |
| 779 | Language models are few-shot learners . In <i>Ad-</i> | |
| 780 | <i>vances in Neural Information Processing Systems</i> , | |
| | volume 33, pages 1877–1901. Curran Associates, | 781 |
| | Inc. | 782 |
| | Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, | |
| | Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. | 783 |
| | 2023. Chateval: Towards better llm-based evaluators | 784 |
| | through multi-agent debate . | 785 |
| | | 786 |
| | Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg | |
| | Durrett. 2022. Generating literal and implied sub- | 787 |
| | questions to fact-check complex claims . In <i>Proceed-</i> | 788 |
| | <i>ings of the 2022 Conference on Empirical Methods</i> | 789 |
| | <i>in Natural Language Processing</i> , pages 3495–3516, | 790 |
| | Abu Dhabi, United Arab Emirates. Association for | 791 |
| | Computational Linguistics. | 792 |
| | | 793 |
| | Wenhu Chen. 2023. Large language models are few(1)- | 794 |
| | shot table reasoners . In <i>Findings of the Associa-</i> | 795 |
| | <i>tion for Computational Linguistics: EACL 2023</i> , | 796 |
| | pages 1120–1130, Dubrovnik, Croatia. Association | 797 |
| | for Computational Linguistics. | 798 |
| | Cheng-Han Chiang and Hung-yi Lee. 2023. Can large | 799 |
| | language models be an alternative to human evalua- | 800 |
| | tions? In <i>Proceedings of the 61st Annual Meeting of</i> | 801 |
| | <i>the Association for Computational Linguistics (Vol-</i> | 802 |
| | <i>ume 1: Long Papers)</i> , pages 15607–15631, Toronto, | 803 |
| | Canada. Association for Computational Linguistics. | 804 |
| | Leon Derczynski, Kalina Bontcheva, Maria Liakata, | 805 |
| | Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz | 806 |
| | Zubiaga. 2017. SemEval-2017 task 8: RumourEval: | 807 |
| | Determining rumour veracity and support for ru- | 808 |
| | mours . In <i>Proceedings of the 11th International</i> | 809 |
| | <i>Workshop on Semantic Evaluation (SemEval-2017)</i> , | 810 |
| | pages 69–76, Vancouver, Canada. Association for | 811 |
| | Computational Linguistics. | 812 |
| | John Dougrez-Lewis, Elena Kochkina, Miguel Arana- | 813 |
| | Catania, Maria Liakata, and Yulan He. 2022. PHE- | 814 |
| | MEPlus: Enriching social media rumour verifica- | 815 |
| | tion with external evidence . In <i>Proceedings of the</i> | 816 |
| | <i>Fifth Fact Extraction and VERification Workshop</i> | 817 |
| | <i>(FEVER)</i> , pages 49–58, Dublin, Ireland. Association | 818 |
| | for Computational Linguistics. | 819 |
| | John Dougrez-Lewis, Maria Liakata, Elena Kochkina, | 820 |
| | and Yulan He. 2021. Learning disentangled latent | 821 |
| | topics for Twitter rumour veracity classification . In | 822 |
| | <i>Findings of the Association for Computational Lin-</i> | 823 |
| | <i>guistics: ACL-IJCNLP 2021</i> , pages 3902–3908, On- | 824 |
| | line. Association for Computational Linguistics. | 825 |
| | Matthias Fey and Jan E. Lenssen. 2019. Fast graph | 826 |
| | representation learning with PyTorch Geometric. | 827 |
| | In <i>ICLR Workshop on Representation Learning on</i> | 828 |
| | <i>Graphs and Manifolds</i> . | 829 |
| | Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ah- | 830 |
| | met Aker, Arkaitz Zubiaga, Kalina Bontcheva, and | 831 |
| | Leon Derczynski. 2019. SemEval-2019 task 7: Ru- | 832 |
| | mourEval, determining rumour veracity and support | 833 |
| | for rumours . In <i>Proceedings of the 13th International</i> | 834 |
| | <i>Workshop on Semantic Evaluation</i> , pages 845–854, | 835 |
| | Minneapolis, Minnesota, USA. Association for Com- | 836 |
| | putational Linguistics. | 837 |

| | | |
|-----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 838 | 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? | 895 |
| 839 | | 896 |
| 840 | | 897 |
| 841 | | 898 |
| 842 | | 899 |
| 843 | Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs . In <i>Advances in Neural Information Processing Systems</i> , volume 30. Curran Associates, Inc. | 900 |
| 844 | | 901 |
| 845 | | 902 |
| 846 | | 903 |
| 847 | Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification . In <i>Findings of the Association for Computational Linguistics: NAACL 2022</i> , pages 1259–1277, Seattle, United States. Association for Computational Linguistics. | 904 |
| 848 | | 905 |
| 849 | | 906 |
| 850 | | 907 |
| 851 | | 908 |
| 852 | | 909 |
| 853 | | 910 |
| 854 | Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 4198–4205, Online. Association for Computational Linguistics. | 911 |
| 855 | | 912 |
| 856 | | 913 |
| 857 | | 914 |
| 858 | | 915 |
| 859 | | 916 |
| 860 | Hao Jia, Honglei Wang, and Xiaoping Zhang. 2022. Early detection of rumors based on source tweet-word graph attention networks . <i>PLoS One</i> , 17(7):e0271224. | 917 |
| 861 | | 918 |
| 862 | | 919 |
| 863 | | 920 |
| 864 | Shailza Jolly, Pepa Atanasova, and Isabelle Augenstein. 2022. Generating fluent fact checking explanations with unsupervised post-editing . <i>Information</i> , 13(10). | 921 |
| 865 | | 922 |
| 866 | | 923 |
| 867 | Ashkan Kazemi, Zehua Li, Verónica Pérez-Rosas, and Rada Mihalcea. 2021. Extractive and abstractive explanations for fact-checking and evaluation of news . In <i>Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda</i> , pages 45–50, Online. Association for Computational Linguistics. | 924 |
| 868 | | 925 |
| 869 | | 926 |
| 870 | | 927 |
| 871 | | 928 |
| 872 | | 929 |
| 873 | | 930 |
| 874 | Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization . In <i>3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings</i> . | 931 |
| 875 | | 932 |
| 876 | | 933 |
| 877 | | 934 |
| 878 | | 935 |
| 879 | 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 . <i>Information Processing and Management</i> , 60(1):103116. | 936 |
| 880 | | 937 |
| 881 | | 938 |
| 882 | | 939 |
| 883 | | 940 |
| 884 | | 941 |
| 885 | Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. All-in-one: Multi-task learning for rumour verification . In <i>Proceedings of the 27th International Conference on Computational Linguistics</i> , pages 3402–3413, Santa Fe, New Mexico, USA. Association for Computational Linguistics. | 942 |
| 886 | | 943 |
| 887 | | 944 |
| 888 | | 945 |
| 889 | | 946 |
| 890 | | 947 |
| 891 | Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners . In <i>Advances in Neural Information Processing Systems</i> . | 948 |
| 892 | | 949 |
| 893 | | 950 |
| 894 | | 950 |
| | 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 <i>Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)</i> , pages 21–30, Dominican Republic. Association for Computational Linguistics. | |
| | Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 7740–7754, Online. Association for Computational Linguistics. | |
| | Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProofVer: Natural logic theorem proving for fact verification . <i>Transactions of the Association for Computational Linguistics</i> , 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 <i>Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)</i> , 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 <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 7871–7880, Online. Association for Computational Linguistics. | |
| | Zachary C. Lipton. 2018. The mythos of model interpretability . <i>Commun. ACM</i> , 61(10):36–43. | |
| | Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. 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 . <i>CoRR</i> , abs/1907.11692. | |
| | Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023b. Calibrating llm-based evaluator . | |
| | Mamta and Asif Ekbal. 2022. Adversarial sample generation for aspect based sentiment classification . In <i>Findings of the Association for Computational</i> | |

| | | | |
|------|-------------------------------------------------------------------------------|--------------------------------------------------------------------------------|------|
| 951 | <i>Linguistics: ACL-IJCNLP 2022</i> , pages 478–492, | Avanti Shrikumar, Peyton Greenside, and Anshul Kun- | 1006 |
| 952 | Online only. Association for Computational Linguistics. | daje. 2017. Learning important features through | 1007 |
| 953 | | propagating activation differences. In <i>International</i> | 1008 |
| 954 | Edoardo Mosca, Maximilian Wich, and Georg Groh. | <i>conference on machine learning</i> , pages 3145–3153. | 1009 |
| 955 | 2021. Understanding and interpreting the impact of | PMLR. | 1010 |
| 956 | user context in hate speech detection . In <i>Proceedings</i> | | |
| 957 | <i>of the Ninth International Workshop on Natural Lan-</i> | Dominik Stambach and Elliott Ash. 2020. e-fever: Ex- | 1011 |
| 958 | <i>guage Processing for Social Media</i> , pages 91–102, | planations and summaries for automated fact check- | 1012 |
| 959 | Online. Association for Computational Linguistics. | ing. In <i>Conference for Truth and Trust Online</i> . | 1013 |
| 960 | Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, | | |
| 961 | Caglar Gulcehre, and Bing Xiang. 2016. Abstrac- | Yingcheng Sun and Kenneth Loparo. 2019. Topic shift | 1014 |
| 962 | tive text summarization using sequence-to-sequence | detection in online discussions using structural con- | 1015 |
| 963 | RNNs and beyond . In <i>Proceedings of the 20th</i> | text . In <i>2019 IEEE 43rd Annual Computer Software</i> | 1016 |
| 964 | <i>SIGNLL Conference on Computational Natural Lan-</i> | <i>and Applications Conference (COMPSAC)</i> , volume 1, | 1017 |
| 965 | <i>guage Learning</i> , pages 280–290, Berlin, Germany. | pages 948–949. | 1018 |
| 966 | Association for Computational Linguistics. | | |
| 967 | Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan | Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. | 1019 |
| 968 | Luu, William Yang Wang, Min-Yen Kan, and Preslav | Axiomatic attribution for deep networks. In <i>Proceed-</i> | 1020 |
| 969 | Nakov. 2023. Fact-checking complex claims with | <i>ings of the 34th International Conference on Machine</i> | 1021 |
| 970 | program-guided reasoning . In <i>Proceedings of the</i> | <i>Learning - Volume 70, ICML'17</i> , page 3319–3328. | 1022 |
| 971 | <i>61st Annual Meeting of the Association for Compu-</i> | JMLR.org. | 1023 |
| 972 | <i>tational Linguistics (Volume 1: Long Papers)</i> , pages | | |
| 973 | 6981–7004, Toronto, Canada. Association for Com- | James Thorne, Andreas Vlachos, Oana Cocarascu, | 1024 |
| 974 | putational Linguistics. | Christos Christodoulopoulos, and Arpit Mittal. 2018. | 1025 |
| 975 | Rob Procter, Miguel Arana-Catania, Yulan He, Maria | The fact extraction and VERification (FEVER) | 1026 |
| 976 | Liakata, Arkaitz Zubiaga, Elena Kochkina, and | shared task . In <i>Proceedings of the First Workshop on</i> | 1027 |
| 977 | Runcong Zhao. 2023. Some observations on fact- | <i>Fact Extraction and VERification (FEVER)</i> , pages 1– | 1028 |
| 978 | checking work with implications for computational | 9, Brussels, Belgium. Association for Computational | 1029 |
| 979 | support . | Linguistics. | 1030 |
| 980 | Rob Procter, Farida Vis, and Alex Voss. 2013. Reading | Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob | 1031 |
| 981 | the riots on twitter: methodological innovation for | Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz | 1032 |
| 982 | the analysis of big data . <i>International Journal of</i> | Kaiser, and Illia Polosukhin. 2017. Attention is all | 1033 |
| 983 | <i>Social Research Methodology</i> , 16(3):197–214. | you need . In <i>Advances in Neural Information Pro-</i> | 1034 |
| 984 | Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, | <i>cessing Systems</i> , volume 30. Curran Associates, Inc. | 1035 |
| 985 | Livio Baldini Soares, Michael Collins, Zachary C. | Petar Veličković, Guillem Cucurull, Arantxa Casanova, | 1036 |
| 986 | Lipton, Graham Neubig, and William W. Cohen. | Adriana Romero, Pietro Lio, and Yoshua Bengio. | 1037 |
| 987 | 2022. Evaluating Explanations: How Much Do Ex- | 2018. Graph attention networks. <i>International Con-</i> | 1038 |
| 988 | planations from the Teacher Aid Students? <i>Transac-</i> | <i>ference on Learning Representations</i> . | 1039 |
| 989 | <i>tions of the Association for Computational Linguis-</i> | Erik Štrumbelj and Igor Kononenko. 2014. Explaining | 1040 |
| 990 | <i>tics</i> , 10:359–375. | prediction models and individual predictions with fea- | 1041 |
| 991 | Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: | ture contributions . <i>Knowl. Inf. Syst.</i> , 41(3):647–665. | 1042 |
| 992 | Sentence embeddings using siamese bert-networks . | Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, | 1043 |
| 993 | In <i>Proceedings of the 2019 Conference on Empirical</i> | and Xuanjing Huang. 2020. Heterogeneous graph | 1044 |
| 994 | <i>Methods in Natural Language Processing</i> . Associa- | neural networks for extractive document summariza- | 1045 |
| 995 | tion for Computational Linguistics. | tion . In <i>Proceedings of the 58th Annual Meeting of</i> | 1046 |
| 996 | Daniel Russo, Serra Sinem Tekiroğlu, and Marco | <i>the Association for Computational Linguistics</i> , pages | 1047 |
| 997 | Guerini. 2023. Benchmarking the generation of fact | 6209–6219, Online. Association for Computational | 1048 |
| 998 | checking explanations . <i>Transactions of the Associa-</i> | Linguistics. | 1049 |
| 999 | <i>tion for Computational Linguistics</i> , 11:1250–1264. | Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui | 1050 |
| 1000 | Michael Schlichtkrull, Zhijiang Guo, and Andreas Vla- | Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, | 1051 |
| 1001 | chos. 2023. Averitec: A dataset for real-world claim | and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? | 1052 |
| 1002 | verification with evidence from the web . | a preliminary study . | 1053 |
| 1003 | Chenhui Shen, Liying Cheng, Yang You, and Lidong | Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, | 1054 |
| 1004 | Bing. 2023. Are large language models good evaluat- | Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and | 1055 |
| 1005 | tors for abstractive summarization? | Zhifang Sui. 2023b. Large language models are not | 1056 |
| | | fair evaluators . | 1057 |
| | | Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita | 1058 |
| | | Bhutani, and Estevam Hruschka. 2023. Less is more | 1059 |
| | | for long document summary evaluation by llms . | 1060 |

1061 Jianfei Yu, Jing Jiang, Ling Min Serena Khoo,
1062 Hai Leong Chieu, and Rui Xia. 2020. [Coupled hi-](#)
1063 [erarchical transformer for stance-aware rumor veri-](#)
1064 [fication in social media conversations](#). In *Proceed-*
1065 [ings of the 2020 Conference on Empirical Methods](#)
1066 [in Natural Language Processing \(EMNLP\)](#), pages
1067 1392–1401, Online. Association for Computational
1068 Linguistics.

1069 Xiaohui Zhang, Qianzhou Du, and Zhongju Zhang.
1070 2020. An explainable machine learning framework
1071 for fake financial news detection. In *Proceedings*
1072 [of the 41st International Conference on Information](#)
1073 [Systems, ICIS 2020, Making Digital Inclusive: Blend-](#)
1074 [ing the Local and the Global, Hyderabad, India,](#)
1075 [December 13-16, 2020](#). Association for Information
1076 Systems.

1077 Xinpeng Zhang, Shuzhi Gong, and Richard O. Sinnott.
1078 2021a. [Social media rumour detection through graph](#)
1079 [attention networks](#). In *2021 IEEE Asia-Pacific Con-*
1080 [ference on Computer Science and Data Engineering](#)
1081 [\(CSDE\)](#), pages 1–6.

1082 Zijian Zhang, Koustav Rudra, and Avishek Anand.
1083 2021b. [Faxplainac: A fact-checking tool based on ex-](#)
1084 [plainable models with human correction in the loop](#).
1085 [CIKM '21](#), page 4823–4827, New York, NY, USA.
1086 Association for Computing Machinery.

1087 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan
1088 Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin,
1089 Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang,
1090 Joseph E Gonzalez, and Ion Stoica. 2023. [Judging](#)
1091 [llm-as-a-judge with mt-bench and chatbot arena](#). In
1092 [Advances in Neural Information Processing Systems](#),
1093 volume 36, pages 46595–46623. Curran Associates,
1094 Inc.

1095 Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine
1096 Wong Sak Hoi, and Peter Tolmie. 2016. [Analysing](#)
1097 [how people orient to and spread rumours in social](#)
1098 [media by looking at conversational threads](#). *PLOS*
1099 *ONE*, 11:1–29.

1100 **A** [Guidelines and Examples of for](#)
1101 [Assessing the Informativeness of](#)
1102 [Explanations](#)

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 5: [Example task instructions used in the prompt following a multiple-choice setting.](#)

B Error Analysis of LLM’s performance as Evaluator 1103 1104

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. 1105 1106 1107 1108 1109 1110

We report the performance of ChatGPT, ChatGPT 0614 and GPT-4 as evaluators using the manually annotated set of 200-300 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. 1112 1113 1114 1115 1116 1117 1118

C Pilot Study on Temperature Setting for ChatGPT 1119 1120

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 1121 1122 1123 1124 1125 1126 1127 1128

| |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 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 #sydneyseige |
| 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.

| | | | |
|------|---------------------------------------------------------|--------------------------------------------------------|------|
| 1129 | and temperature values. In particular, we note that | component. The batch size is 20. For the Graph- | 1142 |
| 1130 | when using temperature 0, the evaluations remain | Sage layers, we apply a mean aggregator scheme, | 1143 |
| 1131 | 100% consistent and for non-zero temperature, the | followed by a relu activation. For the Multi-headed | 1144 |
| 1132 | evaluation only impacts the labelling of the last | Attention layer, we use 8 heads. Embeddings gen- | 1145 |
| 1133 | explanation which is less helpful than previous ex- | erated by the "all-MiniLM-L6-v2" model from Sen- | 1146 |
| 1134 | planation candidates. | tence Transformers (Reimers and Gurevych, 2019) | 1147 |
| 1135 | D Experimental Setup | are used to initialise the node representations in the | 1148 |
| 1136 | We train the rumour verification model for 300 | graphs. To avoid overfitting, we randomly dropout | 1149 |
| 1137 | epochs with learning rate 10^{-5} . The training loss | an edge in the graph networks with probability | 1150 |
| 1138 | is cross-entropy. The optimizer algorithm is Adam | 0.1. We use a Nvidia A5000 GPU for our model | 1151 |
| 1139 | (Kingma and Ba, 2015). Hidden channel size is | training. All model implementation is done via | 1152 |
| 1140 | set as 256 for the propagation and dispersion com- | the <i>pytorch-geometric</i> package (Fey and Lenssen, | 1153 |
| 1141 | ponents and 32 hidden channel size for the stance | 2019) for graph neural networks. | 1154 |

| 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**

1155 **E Current Submission colour-coded for**
1156 **the changes we have implemented**
1157 **compared to the previous version of the**
1158 **manuscript**

1159 **Red** stands for removed material and **blue** stands
1160 for new additions.

| 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.