DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection

Herun Wan[*1]   **Shangbin Feng**[*2]   **Zhaoxuan Tan**[3]
Heng Wang[1]   **Yulia Tsvetkov**[2]   **Minnan Luo**[✉1]

[1] School of Computer Science and Technology,
Xi'an Jiaotong University, Xi'an, 710049, China
[2] University of Washington   [3] University of Notre Dame
wanherun@stu.xjtu.edu.cn; shangbin@cs.washington.edu

Abstract

Large language models are limited by challenges in factuality and hallucinations to be directly employed off-the-shelf for judging the veracity of news articles, where factual accuracy is paramount. In this work, we propose DELL that identifies three key stages in misinformation detection where LLMs could be incorporated as part of the pipeline: 1) LLMs could *generate news reactions* to represent diverse perspectives and simulate user-news interaction networks; 2) LLMs could *generate explanations* for proxy tasks (*e.g.*, sentiment, stance) to enrich the contexts of news articles and produce experts specializing in various aspects of news understanding; 3) LLMs could *merge task-specific experts* and provide an overall prediction by incorporating the predictions and confidence scores of varying experts. Extensive experiments on seven datasets with three LLMs demonstrate that DELL outperforms state-of-the-art baselines by up to 16.8% in macro f1-score. Further analysis reveals that the generated reactions and explanations are greatly helpful in misinformation detection, while our proposed LLM-guided merging helps produce better-calibrated predictions. [1]

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities to follow instructions (Ouyang et al., 2022), perform knowledge-intensive tasks (Rubin et al., 2022; Shi et al., 2023), and confront societal challenges (Jiang et al., 2023c; Roy et al., 2023). However, LLMs are also hindered by hallucinations (Kryściński et al., 2020; Pagnoni et al., 2021; Dong et al., 2022), lack of factuality (Kandpal et al., 2023; Mallen et al., 2023), and challenges to adapt to new knowledge (De Cao et al., 2021; Hase et al., 2021). Despite preliminary efforts (Chen and Shu, 2023; Lucas et al., 2023), LLMs cannot yet be employed off-the-shelf for analyzing the veracity of news articles where factual accuracy is paramount (Leite et al., 2023; Hu et al., 2024). Together with emerging risks of generating misinformation at scale (Chen and Shu, 2023; Wu and Hooi, 2023b), these limitations call for new solutions to leverage LLMs to counter online fake news and misinformation campaigns.

While LLMs are not reliable in detecting misinformation directly, we propose **DELL**[2], employing three key stages where LLMs could be integrated to provide more context and explanations for reliable assessment of news veracity (Figure 1):

- Community reactions and comments to news articles have been shown to improve misinformation detection systems (Grover et al., 2022). However, it is not always feasible to collect sufficient volumes of real-time user interactions (He et al., 2023a). Guided by LLMs' potential in simulating human samples and populations (Argyle et al., 2023), we employ LLMs to **generate synthetic reactions and comments** to news articles from diverse perspectives, turning the news context into a rich network of user-news interactions.

- Previous research shows that additional pragmatic contexts such as sentiment and stance, as well as external knowledge help aid misinformation detection (Zhang et al., 2021; Hu et al., 2021; Sengan et al., 2023). To this end, we employ LLMs for **proxy tasks**, *i.e.*, tasks where predictions and explanations could be helpful to better understand the news article. For example, LLMs evaluate the sentiment of news articles and generate both predictions and explanations: these additional contexts are then encoded as initial

---

[2]DELL stands for **D**iverse Reaction Generation; **E**xplainable Proxy Tasks; and **LL**M-Based Expert Ensemble.
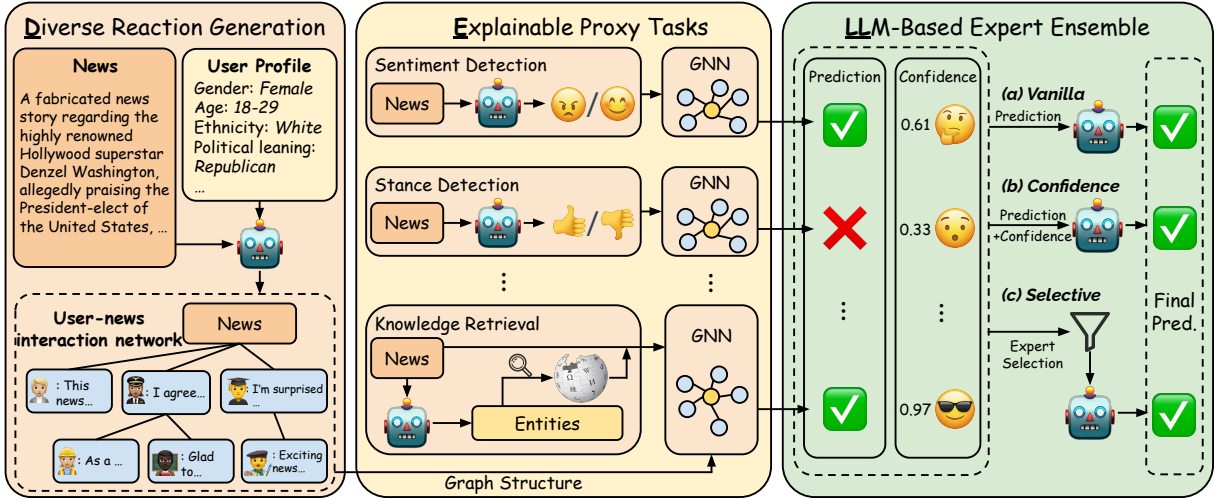
Figure 1: Overview of DELL. We first employ LLMs to generate news reactions from diverse perspectives and form user-news interaction networks. We then design six explainable proxy tasks to refine the feature embeddings with LLM-generated explanations. We finally propose three LLM-based strategies to selectively merge the predictions of task-specific experts and enhance calibration.

embeddings in the user-news interaction network for classification based on graph neural networks (GNNs). By employing six proxy tasks focusing on the news article and generated reactions, we obtain a suite of specialized predictors that specialize in various aspects of news understanding.

• Finally, we employ LLMs as judges to **merge the task-specific experts** and predict the news veracity. Since not all experts are equally helpful/confident for a given news article, we provide LLMs with the predictions and confidence scores of experts specializing in each proxy task: the LLM is then instructed to selectively incorporate the predictions of experts for an overall decision.

We conduct extensive experiments to evaluate DELL and state-of-the-art baselines with three LLMs on seven datasets spanning three tasks related to news veracity, featuring both human-written and machine-generated misinformation. DELL outperforms the strongest baseline across all datasets, achieving an improvement of up to 16.8% in macro f1-score. Further analysis reveals that LLM-generated news reactions and explanations to proxy tasks contribute greatly to model performance, while the LLM-guided expert merging results in better-calibrated misinformation detectors for both human- and machine-generated news.

2 Methodology

We propose three strategies to integrate LLMs in evaluating news veracity: (i) **Diverse Reaction**

Generation, leveraging LLMs to generate synthetic news reactions from diverse perspectives and forming networks of user-news interactions; (ii) **Explainable Proxy Tasks**, enriching news contexts and refining node embeddings in user-news interaction networks with LLM-generated task explanations; (iii) **LLM-Based Expert Ensemble**, adopting LLMs to selectively merge the predictions of task-specific experts and enhance calibration.

2.1 Diverse Reaction Generation

Integrating the public discourse to evaluate news veracity is widely employed to better ground news articles and provide more context (Grover et al., 2022; Sheng et al., 2022; Wu and Hooi, 2023a; Shovon and Shin, 2023). However, real-world comments and reactions are challenging to collect, while malicious comments aiming to bolster misinformation might be removed from social media platforms and hinder reproducible research (Jung et al., 2020; Grover et al., 2022; He et al., 2023a). Motivated by LLMs' successes in simulating human samples (Argyle et al., 2023) and reflecting diverse perspectives (Sorensen et al., 2024), we propose to generate synthetic comments and reactions by LLMs, simulating how populations from diverse perspectives might respond to news articles.

Diverse User Attribute We first define the space of social media user attributes to simulate. Specifically, each synthetic user is represented as an intersection of seven categories:[3] *gender*, *age*, *ethnicity*,

---
[3]We select these categories and attributes from The Pew Research Center's American Trends Panel. Full list of poten-

education, *family income*, *political leaning*, and *voter registration*. Formally, for a user attribute P_i ($1 \leq i \leq n, n = 7$), its candidate set is $\{p_i^j\}_{j=1}^{n_i}$ where n_i denote the number of possibilities for a given attribute category. We sample uniformly for each user attribute to represent a social media user. We then verbalize these attributes and concatenate them as the prompt u for the synthetic user.

Generating User-News Networks Aside from news content, the non-sequential propagation structure of news comments is shown to aid in evaluating news veracity (Ma et al., 2018; Lu and Li, 2020; Ma et al., 2023a). Formally, given a news article s, we aim to generate a user-news interaction network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the node and edge sets. We develop three strategies for LLMs to simulate the comment propagation process: (i) generate a comment on the news article; (ii) generate a comment on a given comment; and (iii) select a comment to engage:

- **Comment on news**. We first generate a synthetic user description u (§2.1) and append the following prompt: "*You view a piece of news with the following content. News: s*". The LLM is then instructed to generate a comment representing the user's perspective, specifically with the prompt "*Please comment on this news on social media.*"

- **Comment on a comment**. Similarly, we first provide LLMs with the user description u and news article s. We append a comment chain $C = [c_1 \| c_2 \| \dots \| c_m]$, where c_i is a comment on c_{i-1}. The LLM is then instructed to generate a comment to the last comment with "*Please reply to the last comment.*"

- **Select a comment to comment**. Social media users would selectively engage with certain comments informed by their perspectives. We employ LLMs to simulate this process by appending u, s, and multiple comment chains C, while instructing the LLM with "*Please select a comment chain that you would most like to reply.*"

We iteratively adopt these prompts to generate a user-news interaction network for a given news article. Algorithm 1 in Appendix A.2 presents details on the user-news network generation process.

tial attributes and example prompts in Appendix A.1.

2.2 Explainable Proxy Tasks

Integrating LLM-generated contexts about a given document has proven effective in analyzing text-attribute graphs such as scholarly networks (He et al., 2023b; Chen et al., 2023c; Li et al., 2023a). In the domain of misinformation detection, there is often much implied context that goes beyond the news text itself, such as author stances, sentiment, external knowledge, and more. We propose to employ LLM-generated explanations for proxy tasks, *i.e.*, tasks that help evaluate news veracity, enriching news contexts and refining the feature embeddings of user-news interaction networks with the generated explanations. Specifically, we propose four proxy tasks to enhance news articles:

- **Sentiment Analysis** News articles often feature sentiment signals that are indicative of their veracity (Zhang et al., 2021). We employ six basic emotions (Ekman et al., 1999) (*e.g.*, *anger* and *surprise*) and prompt LLMs to choose the three most likely emotions and provide explanations.

- **Framing Detection** Framing is a strategic device in political communication (Entman, 1993) and has been an integral part of evaluating news veracity (Kwak et al., 2020; Mendelsohn et al., 2021). Similarly, we follow the taxonomy of 14 media frames (Card et al., 2015a) (*e.g.*, *economic*) and prompt LLMs to choose the five most likely media frames and provide explanations.

- **Propaganda Tactics Detection** Propaganda tactics are employed to influence people's mindsets to advance a specific agenda (Glowacki et al.). We follow the taxonomy of 19 propaganda tactics (Piskorski et al., 2023) (*e.g.*, *doubt* and *red herring*) and employ LLMs to identify the underlying tactics in news articles with explanations.

- **Knowledge Retrieval** Retrieval-augmented language models (Borgeaud et al., 2022; Shi et al., 2023; Asai et al., 2023; Chen et al., 2023b) have demonstrated impressive potential to expand the knowledge access of LLMs. We employ LLMs to identify key entities in a news article through prompting and retrieve Wikipedia passages about these entities[4]. We prepend the retrieved external knowledge in the news article to facilitate better contextual understanding.

Besides news content, we also propose two proxy tasks to enhance the generated comments:

---

[4] We employ the Wikipedia API for retrieval.

- **Stance Detection** Given two text nodes s_1 and s_2 (news or comments) that are connected in the user-news interaction network $\mathcal{G}$, we employ LLMs to evaluate whether s_1 and s_2 are *supportive*, *neutral*, or *opposed* to each other with explanations.

- **Response Characterization** Given two text nodes s_1 and s_2 (news or comments) in $\mathcal{G}$, we employ LLMs to analyze whether one is in response to another. The generated explanations would help better understand the propagation structure of news and comments.

By employing any of the six proxy tasks[5], we obtain an LLM-generated explanation paragraph s_{ext} that analyzes the news article from one specialized aspect. We leverage the LLM-generated explanations to refine the feature embeddings of user-news interaction networks. Specifically, we first adopt a separate encoder-based LM $\mathrm{enc}(\cdot)$ to encode the news article s_{ori} and the explanation s_{ext}, *i.e.*, $\mathbf{h}_{ori(ext)} = \mathrm{enc}(s_{ori(ext)})$, where we employ DeBERTa (He et al., 2021) in practice. We then concatenate $\mathbf{h}_{ori}$ and $\mathbf{h}_{ext}$ and feed it into a linear layer to obtain initial node features $\mathbf{h}^{(0)}$.

We employ graph neural networks as the model for downstream tasks, which conduct message passing over the user-news network. Formally, suppose $\mathbf{h}_{v_i}^{(\ell)}$ is the representation of node v_i at the ℓ-th GNN layer, the feature update procedure is:

$$\mathbf{h}_{v_i}^{(\ell)} = \underset{\forall v_j \in \mathcal{N}(v_i)}{\mathrm{Aggr}} (\{\mathrm{Prop}(\mathbf{h}_{v_i}^{(\ell-1)}; \mathbf{h}_{v_j}^{(\ell-1)})\}),$$

where $\mathcal{N}(v_i)$ denotes the set of neighbors of node v_i, $\mathrm{Aggr}(\cdot)$ and $\mathrm{Prop}(\cdot)$ are aggregation and propagation functions, where GIN (Xu et al., 2019) is employed in practice. To obtain the graph-level representation of $\mathcal{G}$, we employ the mean pooling operator as the $\mathrm{Readout}(\cdot)$ function, *i.e.*,

$$\mathbf{h} = \mathrm{Readout}(\{\mathbf{h}_{v_i}^{(\ell)}\}_{v_i \in \mathcal{V}}).$$

Given a user-news network $\mathcal{G}$ and a label y, we compute the probability of y being the correct prediction as $p(y \mid \mathcal{G}) \propto \exp(\mathrm{MLP}(\mathbf{h}))$, where $\mathrm{MLP}(\cdot)$ denotes an MLP layer. For binary classification, we optimize models using the cross-entropy loss and predict the most plausible label as $\arg\max_y p(y \mid \mathcal{G})$. For multi-label classification, we optimize models using the ZLPR (Su et al., 2022) loss and predict the label set as $\{y : p(y \mid \mathcal{G}) > \lambda\}$, where λ is a hyperparameter.

2.3 LLM-Based Expert Ensemble

By adopting different proxy tasks and LLM-generated explanations, we obtain a set of experts, where each specializes in one proxy task and various aspects of news articles. To obtain an overall prediction, we propose an LLM-based expert ensemble to selectively leverage experts, their predictions, and confidence scores. We first use one sentence d_i to describe each expert, *e.g.*, "This expert focuses on the emotion of news." We then propose three modes for LLMs to merge experts[6]:

Vanilla LLMs are first provided with news content and an instruction, *i.e.*, "*Some experts give predictions about the news.*" We then append the description and prediction of each expert: for an expert e_i with prediction ℓ_i and its description d_i, the expert prompt is "*Expert i: d_i. The expert predicts the label of this news is ℓ_i.*" Finally, the LLM is instructed to reason and generate a final prediction based on the experts' feedback.

Confidence In *Vanilla*, we assume that all experts should be equally important. However, experts could have varying levels of confidence and we take this into account by additionally providing the confidence scores. The confidence scores are obtained from the classification layer of the GNN-based model (§2.2). We aim to improve the calibration of LLM-based expert ensemble by incorporating confidence scores of individual experts.

Selective In *Vanilla* and *Confidence*, we assume that every news article would benefit from the input of all experts. However, this could introduce noise in the LLM reasoning process (Feng et al., 2023d; Zhao et al., 2024). To this end, we propose the *Selective* approach, putting LLMs in charge to selectively activate experts. Specifically, we provide news content and expert descriptions, then prompt LLMs with "*To understand this news, which expert knowledge do you need?*" We ensemble the selected experts with the *Confidence* strategy to obtain the final predictions.

3 Experiment Settings

Models and Settings We leverage *Mistral-7B* (Jiang et al., 2023a), *LLaMA2-70B* (Touvron et al., 2023), and *ChatGPT* as the base LLMs. We mainly employ *Mistral-7B* to generate comments and conduct proxy tasks, and *ChatGPT* to ensemble ex-

---

[5]We provide the prompts for proxy tasks in Appendix A.3.

[6]We provide prompts in Appendix A.4.

| Method | Fake News Detection | | | | Framing Detection | | | | Propaganda Tactic Detection | | | | | |
| | Pheme | | LLM-mis | | MFC | | SemEval-23F | | Generated | | SemEval-20 | | SemEval-23P | |
	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF
ZERO-SHOT	.459	.460	.597	.600	.332	.346	.381	.443	.223	.233	.304	.424	.228	.379
FEW-SHOT	.490	.500	.565	.570	.350	.395	.457	.512	.344	.358	.359	.468	.266	.424
RETRIEVAL	.464	.470	.624	.630	.278	.334	.397	.480	.262	.267	.292	.415	.187	.309
F3 *Z-CoT*	.499	.500	.566	.570	.285	.314	.370	.470	.223	.203	.302	.418	.248	.423
F3 *DeF-Gen*	.410	.410	.477	.480	.319	.354	.381	.468	.284	.290	.331	.508	.259	.396
TAPE W/O GRAPH	.767	.770	.858	.860	.341	.482	.393	.631	.298	.326	.332	.565	.237	.583
DEBERTA	.779	.780	.887	.890	.388	.543	.506	.672	.512	.516	.516	.609	.343	.558
K-HOPS	.374	.490	.421	.470	.332	.407	.362	.466	.206	.193	.350	.448	.280	.393
K-ATTENTION	.325	.450	.407	.450	.348	.418	.413	.496	.214	.211	.310	.409	.198	.318
TAPE W/ GRAPH	.787	.790	.888	.890	.381	.515	.399	.623	.279	.306	.332	.598	.250	.581
GCN	.790	.790	.854	.860	.447	.566	.499	.658	.504	.496	.517	.628	.358	.547
RVNN	.790	.790	.888	.890	.428	.551	.494	.644	.494	.496	.462	.559	.363	.568
DEFEND	.727	.730	.823	.840	.434	.607	.435	.557	.063	.099	.280	.576	.255	.601
HYPHEN	.777	.780	.836	.840	.481	**.634**	.528	.714	.292	.327	.347	.508	.301	.488
GET	.788	.790	.847	.850	.445	.566	.525	.649	.250	.227	.423	.561	.361	.617
WSDMS	.799	.800	.860	.870	.434	.597	.526	.688	.376	.419	.509	.630	.333	.619
DELL *Single*	.810	.810	**.928**	**.930**	.458	.598	.536	.684	.543	.556	.520	.613	.376	.631
DELL *Vanilla*	.810	.810	.926	**.930**	.432	.591	.528	.689	.578	.566	.508	.611	.365	.634
DELL *Confidence*	.810	**.820**	.917	.920	**.509**	.603	**.572**	**.718**	.579	.558	.523	.624	**.386**	.643
DELL *Selective*	**.820**	**.820**	.897	.900	.488	.581	.554	.683	**.598**	**.577**	.525	**.636**	.362	**.652**

Table 1: Performance of DELL and baselines on seven datasets from three misinformation-related tasks. *Single* indicates the best-performing single expert. "MaF" and "MiF" indicates macro- and micro-averaged f1-score. **Bold** indicates the best performance and underline indicates the second best. DELL outperforms state-of-the-art baselines by up to 16.8% in macro f1-score, indicating the success of our LLM integration strategies.

perts. We set the temperature $\tau = 0.6$ for *Mistral-7B* and $\tau = 0.1$ for *ChatGPT*. We present more results from other LLMs in Appendix C.

Baselines We compare DELL with three types of state-of-the-art baselines: 1) **LLM-only**: ZERO-SHOT, FEW-SHOT, RETRIEVAL-augmented generation, F3 *Z-CoT* (Lucas et al., 2023), F3 *DeF-Gen* (Lucas et al., 2023), TAPE W/O GRAPH (He et al., 2023b), and DEBERTA (He et al., 2021); 2) **LLM+Graph**: K-HOPS (Huang et al., 2023a) and K-ATTENTION (Huang et al., 2023a), and TAPE W/ GRAPH (He et al., 2023b); 3) **Graph-based**: GCN (Kipf and Welling, 2017), RVNN (Ma et al., 2018), DEFEND (Shu et al., 2019a), HYPEHN (Grover et al., 2022), GET (Xu et al., 2022), and WSDMS (Yang et al., 2023b). We provide more details about baselines in Appendix B.2.

Tasks and Datasets We evaluate DELL and baselines on three tasks related to chacterizing mis-information, *i.e.*, 1) fake news detection: **Pheme** (Buntain and Golbeck, 2017) and **LLM-mis** (Chen and Shu, 2023), which feature a binary classification setting; 2) framing detection: **MFC** (Card et al., 2015b) and **SemEval-23F** (Piskorski et al., 2023), which feature a multi-label classification setting; 3) propaganda tactic detection: **Gener-**

ated generated by *ChatGPT*, **SemEval-20** (Martino et al., 2020), and **SemEval-23P** (Piskorski et al., 2023), which feature a multi-label classification setting. The datasets are all in English and we provide more dataset details in Appendix B.1. To evaluate the ability to evaluate machine-generated news, **LLM-mis** is an extended version of **FakeNewsNet** (Shu et al., 2020) and **Generated** is generated by LLMs.

4 Results

We present the performance of DELL and state-of-the-art baselines in Table 1. We present more ablation study results in Table 6 in Appendix C.

DELL achieves state-of-the-art performance. DELL outperforms the strongest baseline on all seven benchmarks by 1.46% to 16.80% on macro f1-score, indicating the success of integrating LLMs in multiple stages of news veracity evaluation. We find that LLM-only in-context learning approaches struggle in performance, indicating that LLMs are limited by factuality challenges and hallucinations to evaluate the veracity of news articles.

Generated news reactions help ground news articles. Compared to news-only approaches, models enhanced with generated comments (both ours and

Metric	Real Networks			Simulated Networks							
	Pheme	Twitter15	Twitter16	More	Pheme	LLM-mis	MFC	SemEval-23F	Generated	SemEval-20P	SemEval-23P
betweenness	0.255	0.191	0.234	0.208	0.293	0.291	0.291	0.287	0.291	0.286	0.288
shortest path	2.682	1.904	1.833	2.076	2.925	2.913	2.913	2.869	2.908	2.863	2.879
degree	0.764	0.945	0.962	0.821	0.400	0.399	0.399	0.408	0.402	0.416	0.410
diameter	5.477	2.848	2.605	3.281	6.006	5.942	5.951	5.793	5.929	5.792	5.840

Table 2: The graph indicators of the real and simulated networks. "More" denotes that networks are generated when $\alpha = 0.8$ and $\beta = 0.05$. Our generated networks are statistically similar to those in dataset Pheme as of network structure, indicating our generation strategy could stimulate the network structures similar to the real situation.
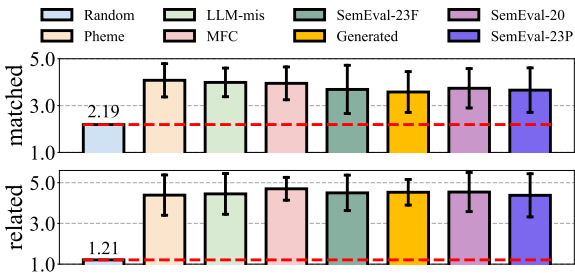

Figure 2: GPT-4 evaluation of whether the LLM-generated comments are related to the news article and match the user attributes, the higher the better from 1 to 5. We present the average value and standard deviation. Compared with randomly paired news ("Random" in the figure), user attributes, and comments, the generated comments generally conform to the user attributes and are relevant to the news articles.
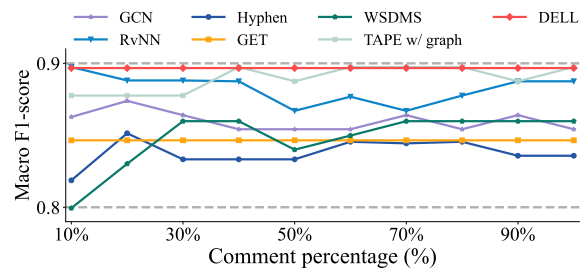

Figure 3: Performance of DELL and baselines on **LLM-mis** when the comments are gradually removed. DELL shows great robustness to the availability of comments.

graph-based baselines) achieve better performance. The average performance on **MFC** of the comment-enhanced models is 15.2% higher on MaF. It indicates that LLM-generated diverse comments are beneficial in characterizing misinformation.

Proxy tasks improve news understanding ability. DELL *single* denotes the performance of the best single expert focusing on one proxy task. We find that a single expert could already achieve a substantial improvement in most cases: for example, on benchmark **Generated**, it achieves a 6.16% improvement on the macro f1-score than the strongest baseline. This indicates that our explainable proxy tasks are effective strategies for incorporating LLMs for evaluating news veracity.

LLMs could ensemble expert predictions. Compared to a single expert, the proposed LLM ensemble strategies achieve improvements on six out of seven datasets. In addition to simple aggregation (*Vanilla*), *Confidence* and *Selective* improve the ensemble by accessing the confidence scores and selectively incorporating certain experts, indicating that LLMs have preliminary capabilities of understanding verbalized confidence scores (Tian et al., 2023; Feng et al., 2024). We further investigate if LLM-based ensembling could lead to better-

calibrated misinformation detectors in Section 5.

5 Analysis

Quality of Generated Comments We verify the quality of LLM-generated comments on whether it matches the user attributes and whether it is related to the news article. We conduct a human evaluation with four annotators to manually evaluate 50 generated comments from two datasets on a five-point Likert scale, where the higher scores mean better quality. The average score is 4.52, the standard deviation is 0.69 and the annotator agreement in Fleiss' Kappa is 0.216, which indicates that annotators generally agree that the LLM-generated comments are related and on-brand for user attributes.

We additionally employ GPT-4 evaluation (Chiang and Lee, 2023; Kim et al., 2023b) for quantitative evaluation, where we randomly sample 700 generated comments and prompt GPT-4 with "*Does the user's comment on the news match the profile?*" and "*Does the comment relate to the news?*" to solicit a response on a five-point Likert scale. Figure 2 demonstrates that the automatic evaluation also finds that the generated comments are consistent with the user attributes and relevant to the news.

We conduct an additional evaluation to "put a more challenging control group of comments generated by the same framework but with a different demographic": we sample comments from users with other attributes, and then employ GPT-4 evaluation to check whether the generated comments
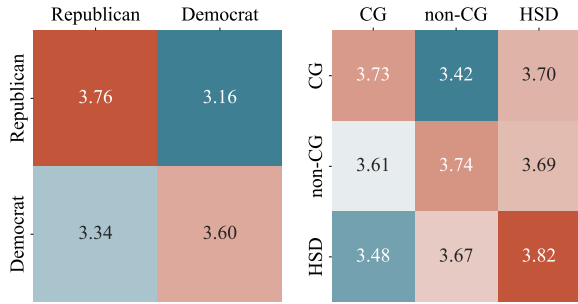
Figure 4: GPT-4 evaluation of the matching degree between different user groups. "CG" denotes "college grad", "non-CG" denotes "haven't graduated from college", and "HSD" denotes "have a high school diploma or less". The diagonal numbers are the highest both row-wise and column-wise, indicating that the generated comments are consistent with the user attributes.
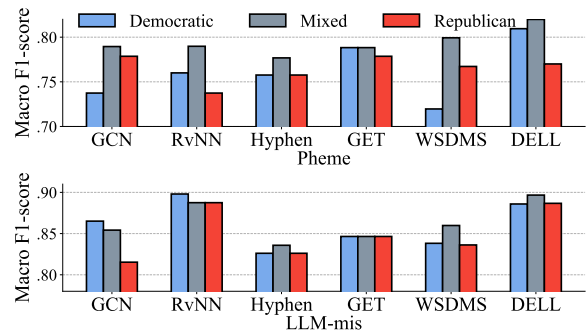

Figure 5: Performance of DELL and baselines when the comments are generated from only one partisan perspective. Models trained on comments from one perspective generally perform worse than diverse comments.

match each attribute. For example, we sample 100 synthetic comments (50 with the attribute Democrat and 50 with the attribute Republican), and then we employ GPT4 to evaluate to what extent, on a scale of 1-5, do these comments match Democrats and Republicans. Then we could obtain 200 scores and draw a heat map. We similarly experiment with the education attribute, spanning "college grad", "haven't graduated from college", and "have a high school diploma or less". In Figure 4, we find that the diagonal numbers, where the user attribute matches what GPT-4 evaluates, are the highest both row-wise and column-wise, indicating that the generated comments are consistent with the user attributes.

Network Generation Ability   To establish that the generated interaction networks resemble real-world networks, we compare our generated networks with the real networks in datasets **Pheme** (Buntain and Golbeck, 2017), **Twitter-15** (Ma et al., 2018), and **Twitter-16** (Ma et al., 2018). Specifically, we calculate the average edge betweenness of each edge, the average shortest path length, the ratio of maximum degree to number of nodes, and the diameter of each graph. Then we average the value over the whole dataset to compare in Table 2. The results show that our generated networks are statistically similar to those in dataset **Pheme** as for network structure, indicating our generation strategy could stimulate the network structures similar to the real situation. In addition, hyperparameters in Algorithm 1 enable the control of generating user reaction networks. For example, by setting $\alpha = 0.8$ and $\beta = 0.05$, generated networks resemble those in datasets **Twitter-15**

and **Twitter-16**. As a result, DELL could reliably simulate real-world user interaction networks and structures through those control measures.

Model Robustness to Comments   Since comments are usually hard to collect and generating comments using LLM could be computationally expensive, detectors should be robust to the amount of comments. We evaluate approaches on the test sets where LLM-generated comments are gradually removed. As demonstrated in Figure 3, DELL drops the least in performance with reduced comments and on dataset **LLM-mis** our performance is almost unchanged. This indicates that DELL benefits greatly from as few as 10% of news comments.

Comment Diversity   We propose to generate diverse comments by employing LLMs to simulate diverse user attributes. To validate this design choice, we re-generate news comments solely with synthetic Republican or Democratic users and evaluate model performance on the fake news detection benchmarks. Figure 5 demonstrates that only considering reactions from a single partisan viewpoint is generally worse, supporting our proposal of integrating diverse comments in fake news detection.

Expert Ablation   Experts are specialized with two types of proxy tasks, focusing on either news content or comments. We conduct ablation studies to examine the impact of different types of proxy tasks. Table 6 demonstrates that: 1) integrating both types of experts leads to better performance, where the performance of a single category drops by up to 15.8%; and 2) experts focusing on proxy tasks of news content generally outperform experts who focus solely on comments, while the two types of proxy tasks are complementary.

Strategy	Variants	Fake News Detection		Framing Detection		Propaganda Tactic Detection		
		Pheme	LLM-mis	MFC	semeval-23F	Generated	semeval-20	semeval-23P
Vanilla	**Original**	.810	.926	.432	.528	.578	.508	.365
	Only Content	.799 (-1.3%)	.885 (-4.4%)	.446 (+3.2%)	.537 (+1.7%)	.570 (-1.3%)	.520 (+2.4%)	.397 (+8.8%)
	Only Comments	.780 (-3.7%)	.927 (+0.1%)	.449 (+4.0%)	.533 (+1.0%)	.436 (-24.6%)	.526 (+3.6%)	.345 (-5.5%)
Confidence	**Original**	.820	.917	.509	.572	.579	.523	.386
	Only Content	.820 (+0.0%)	.907 (-1.1%)	.458 (-9.9%)	.578 (+1.1%)	.556 (-3.9%)	.515 (-1.4%)	.404 (+4.6%)
	Only Comments	.769 (-6.1%)	.907 (-1.0%)	.428 (-15.8%)	.534 (-6.7%)	.548 (-5.4%)	.470 (-10.1%)	.386 (-0.1%)
Select	**Original**	.820	.897	.488	.554	.598	.525	.362
	Only Content	.800 (-2.4%)	.907 (+1.1%)	.477 (-2.2%)	.540 (-2.5%)	.579 (-3.2%)	.526 (+0.1%)	.360 (-0.4%)
	Only Comments	.770 (-6.1%)	.917 (+2.2%)	.426 (-12.7%)	.547 (-1.4%)	.529 (-11.5%)	.507 (-3.4%)	.394 (+8.9%)

Table 3: Ablation study of **expert ensemble**, where only experts of proxy tasks focusing on either news content or comments are retained. We present the macro f1-score for each variant and performance changes compared to the original setup. Diverse experts generally outperform a single type of expert, while experts who focus on news content are generally better than those who focus on comments.
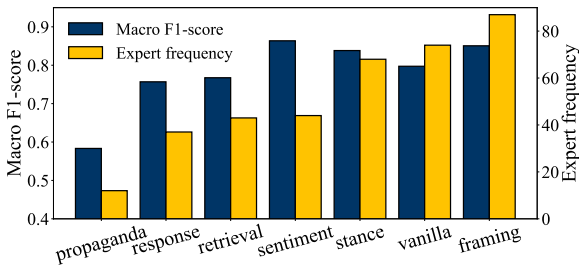

Figure 6: The frequency of expert selections and performance on **Pheme** when a particular expert is selected in the *Selective* approach. Experts who have been selected more times tend to perform better.

Expert Selection   In the *Selective* LLM-based ensemble strategy, LLMs determine which experts are activated and incorporated in the overall decision. To evaluate each expert's contribution, we examine the frequency of expert selection and the performance when a given expert is selected. Figure 6 illustrates that experts who have been selected more times tend to perform better, indicating that LLMs have preliminary capabilities to select helpful experts based on the news content.

Model Calibration   Robust fake news detectors should provide not only a binary prediction but also a well-calibrated confidence score to facilitate content moderation. We evaluate how well DELL and baselines are calibrated with the fake news detection datasets in Figure 8. We use the probability of the prediction token ("fake" or "real") from the LLM as the confidence score, bin it into five buckets (0.5 to 1.0), and calculate the estimated calibration error (ECE) (Guo et al., 2017). It is demonstrated that DELL are better-calibrated with an ECE of 0.2357 while achieving an improvement of up to 19.1% compared to baselines. We hypothesize that by integrating expert confidence scores in the LLM-guided ensemble, the overall decision
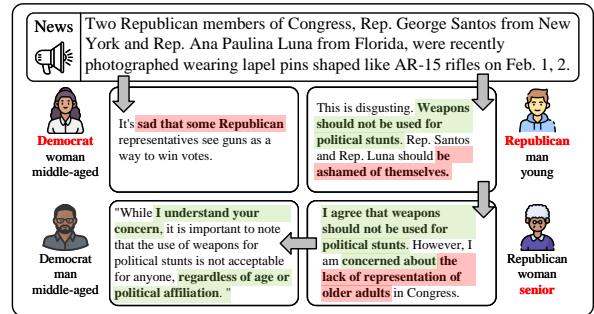

Figure 7: An example of generated comments related to a news article. The red areas indicate the comments match the user profiles, while the green areas indicate strong continuity in the comments of the comment chain. This case highlights the quality and contribution of the LLM-generated comments.

is better-calibrated and thus more trustworthy.

Case Study   We study a specific case of news article and its LLM-generated comments in Figure 7. The red area indicates that the generated comments match the user attributes about partisanship and age groups. The green areas indicate strong continuity in the comment chain. Overall, the example showcases the effectiveness of DELL in generating diverse comments that ground news articles and facilitate characterization.

6   Related Work

Existing fake news detection methods (Zeng and Gao, 2022; Biamby et al., 2022; Mendes et al., 2023; Sung et al., 2023; Xu et al., 2023a; Liao et al., 2023) mostly fall into text-based (Pelrine et al., 2021; Jin et al., 2022; Chen et al., 2023d) and graph-based approaches (Wu et al., 2022; Zhou et al., 2022; Karami et al., 2023; Feng et al., 2023e; Lin et al., 2023; Phan et al., 2023; Chang et al., 2023; Ma et al., 2023b).   Text-only approaches take news context and employ NLP methodologies
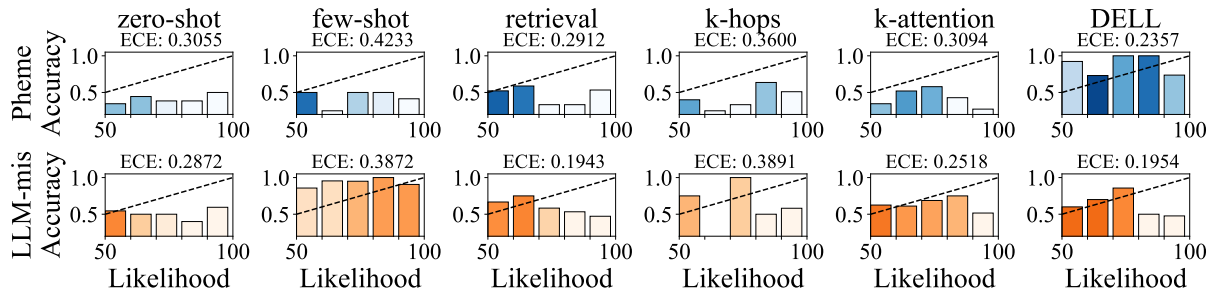
Figure 8: Calibration of DELL and baselines in the fake news detection benchmarks. ECE denotes estimated calibration error, the lower the better. The dashed line indicates perfect calibration, while the bar's color is darker when it is closer to perfect calibration. DELL achieves low ECE scores and thus is better-calibrated.

for classification such as recurrent neural networks (Goonathilake and Kumara, 2020; Liu et al., 2023), attention mechanism (Shu et al., 2019a; Dun et al., 2021), and pre-trained language models (Hartl and Kruschwitz, 2022). In addition to solely considering news content, graph-based approaches first construct networks composed of entities such as news articles, sources (Nguyen et al., 2020), users (Shu et al., 2019b; Dou et al., 2021), and more. These approaches then employ graph neural networks (Bian et al., 2020; Zhang et al., 2024) for classification. Among graph-based approaches, the widely used is to employ comments, *i.e.*, user reactions to news article on social media (Yang et al., 2021; Tian et al., 2022; Mehta et al., 2022; Yang et al., 2023b; Russo et al., 2023; Min and Ananiadou, 2023). In this work, we seek to employ LLMs to generate synthetic comments from diverse perspectives to complement the scarce and incomplete comment networks in real-world datasets (Jung et al., 2020; Micallef et al., 2020; Heidari et al., 2021).

With the advent of autoregressive large language models, previous works have attempted to gauge their risks and generate misinformation with LLMs (Zellers et al., 2019; Fung et al., 2021; Huang et al., 2023c; Wang et al., 2023). They find that LLMs are capable of generating misinformation that is challenging to detect and characterize (Huang et al., 2023b; Chen and Shu, 2023; Pan et al., 2023b; Goldstein et al., 2023; Su et al., 2023b; Xu et al., 2023b; Uchendu et al., 2023). On the other hand, researchers have attempted to employ LLMs off-the-shelf for misinformation research through prompting and in-context learning (Stiff and Johansson, 2022; Gabriel et al., 2022; Kim et al., 2023a; Pelrine et al., 2023; Russo et al., 2023; Jiang et al., 2023b; Nakshatri et al., 2023; Sundriyal et al., 2023; Su et al., 2023a; Li et al., 2023b; Chen et al., 2023a; Feng et al., 2023c; Yue et al., 2023; Yang et al., 2023a; Choi and Ferrara, 2024). We argue

that LLMs face challenges of hallucination (Ji et al., 2023; Du et al., 2023), factuality (Kandpal et al., 2023; Pan et al., 2023a), and temporal knowledge update (Feng et al., 2023a; Luo et al., 2024): as a result, they could not be directly used off-the-shelf for predicting a True-of-False label since they lack accurate and up-to-date information about real-world news events, while such information is crucial in characterizing fake news campaigns. To this end, we identify three key stages in evaluating news veracity and propose strategies to integrate LLMs in countering online misinformation campaigns.

7 Conclusion

We propose DELL for identifying fake news where LLMs could be incorporated as part of the pipeline. First, we employ LLMs to generate news reactions from diverse perspectives and simulate user-news networks. Second, we design six explainable proxy tasks that help identify misinformation. LLMs perform these tasks and generate explanations to produce experts specializing in various aspects of news articles. Finally, we develop three strategies for LLMs to merge task-specific experts and provide an overall prediction. Extensive experiments demonstrate that DELL achieves state-of-the-art performance on three tasks across seven datasets, presenting a misinformation detector better calibrated and better grounded in diverse perspectives.

academy of engineering "The Online and Offline Mixed Educational Service System for 'The Belt and Road' Training in MOOC China". We would like to express our gratitude for the support of K. C. Wong Education Foundation.

Limitation

While DELL could generate synthetic news reactions from diverse perspectives and form networks of user-news interactions, the iterative process with LLMs in computationally heavy. Scaling our solution to the real-world scale of millions of real-time news reactions could be challenging, while we expect efficient LLM inference approaches could help alleviate this limitation.

While we develop six proxy tasks for LLMs to generate explanations and enrich news contexts, they may not be able to fully tap into the diverse capabilities of LLMs and their potential for evaluating the veracity of news articles. Future work could focus on automatically generating and proposing proxy tasks for a more general LLM-as-enhancer framework.

Ethics Statement

The development of fake news detectors is essential in countering online misinformation campaigns. This research demonstrates that LLMs could be integrated as part of the news analysis pipeline. However, it may increase the risk of dual-use, where malicious actors may develop advanced misinformation campaigns that are evasive to LLM-generated comments and explanations. We will establish controlled access to ensure that the data and trained model checkpoint are only publicly available to researchers.

LLMs have been widely shown to have inherent social biases (Bender et al., 2021; Jin et al., 2021; Shaikh et al., 2023), and such biases could have an impact on fake news detection (Feng et al., 2023b). Informed by LLMs' internal biases, stereotypes, and spurious correlations, DELL might struggle to simulate certain demographic groups and provide incorrect explanations of news articles. We argue that the predictions of DELL should be interpreted as an initial screening, while content moderation decisions should be made with experts in the loop.

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Giscard Biamby, Grace Luo, Trevor Darrell, and Anna Rohrbach. 2022. Twitter-comms: Detecting climate, covid, and military multimodal misinformation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1530–1549.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Cody Buntain and Jennifer Golbeck. 2017. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015a. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015b. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*.

Yi-Ting Chang, Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. 2023. Beyond detection: A defend-and-summarize strategy for robust and interpretable rumor analysis on social media. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11538–11556.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Mengyang Chen, Lingwei Wei, Han Cao, Wei Zhou, and Songlin Hu. 2023a. Can large language models understand content and propagation for misinformation detection: An empirical study. *arXiv preprint arXiv:2311.12699*.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Dong Yu, and Hongming Zhang. 2023b. Dense x retrieval: What retrieval granularity should we use? *arXiv preprint arXiv:2312.06648*.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. 2023c. Exploring the potential of large language models (llms)in learning on graphs. *SIGKDD Explor.*, 25(2):42–61.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023d. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

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

Eun Cheol Choi and Emilio Ferrara. 2024. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1441–1449.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2022. A survey of natural language generation. *ACM Computing Surveys*, 55(8):1–38.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055.

Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. 2023. Quantifying and attributing the hallucination of large language models via association analysis. *arXiv preprint arXiv:2309.05217*.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89.

Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Robert Entman. 1993. Framing: Toward clarification of a fractured paradigm. *The Journal of Communication*, 43:51–58.

Chao Feng, Xinyu Zhang, and Zichu Fei. 2023a. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023b. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11737–11762. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023c. Cook: Empowering general-purpose language models with modular and collaborative knowledge. *arXiv preprint arXiv:2305.09955*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023d. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Shangbin Feng, Zhaoxuan Tan, Wenqian Zhang, Zhenyu Lei, and Yulia Tsvetkov. 2023e. KALM: knowledge-aware integration of local, document, and global contexts for long document understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2116–2138. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Monika Glowacki, Vidya Narayanan, Sam Maynard, Gustavo Hirsch, Bence Kollanyi, Lisa-Maria Neudert, Phil Howard, Thomas Lederer, and Vlad Barash. News and political information consumption in mexico: Mapping the 2018 mexican presidential election on twitter and facebook.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

MDP P Goonathilake and PPN V Kumara. 2020. Cnn, rnn-lstm based hybrid approach to detect state-of-the-art stance-based fake news on social media. In *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 23–28. IEEE.

Karish Grover, SM Angara, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Public wisdom matters! discourse-aware hyperbolic fourier co-attention for social text classification. *Advances in Neural Information Processing Systems*, 35:9417–9431.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Philipp Hartl and Udo Kruschwitz. 2022. Applying automatic text summarization for fake news detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2702–2713.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023a. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2023b. Harnessing explanations: Llm-to-lm interpreter for enhanced text-attributed graph representation learning.

In *The Twelfth International Conference on Learning Representations*.

Maryam Heidari, Samira Zad, Parisa Hajibabaee, Masoud Malekzadeh, SeyyedPooya HekmatiAthar, Ozlem Uzuner, and James H Jones. 2021. Bert model for fake news detection based on social bot activities in the covid-19 pandemic. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pages 0103–0109. IEEE.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

Jin Huang, Xingjian Zhang, Qiaozhu Mei, and Jiaqi Ma. 2023a. Can llms effectively leverage graph structural information: When and why. In *NeurIPS 2023 Workshop: New Frontiers in Graph Learning*.

Kung-Hsiang Huang, Kathleen Mckeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023b. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589.

Kung-Hsiang Huang, Kathleen R. McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023c. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14571–14589. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023b. Disinformation detection: An evolving challenge in the age of llms. *arXiv preprint arXiv:2309.15847*.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tanga, Haizhou Wang, and Wenxian Wang. 2023c. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *arXiv preprint arXiv:2310.05650*.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5746–5754.

Anna-Katharina Jung, Björn Ross, and Stefan Stieglitz. 2020. Caution: Rumors ahead—a case study on the debunking of false information on twitter. *Big Data & Society*, 7(2):2053951720980127.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Mansooreh Karami, David Mosallanezhad, Paras Sheth, and Huan Liu. 2023. Silence speaks volumes: Re-weighting techniques for under-represented users in fake news detection. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1430–1437. IEEE.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. 2023a. Covid-19 vaccine misinformation in middle income countries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3903–3915.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023b. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th ACM Conference on Web Science*, pages 305–314.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2023. Detecting misinformation with llm-predicted credibility signals and weak supervision. *arXiv preprint arXiv:2309.07601*.

Yuhan Li, Zhixun Li, Peisong Wang, Jia Li, Xiangguo Sun, Hong Cheng, and Jeffrey Xu Yu. 2023a. A survey of graph meets large language model: Progress and future directions. *arXiv preprint arXiv:2311.12399*.

Zizhong Li, Haopeng Zhang, and Jiawei Zhang. 2023b. A revisit of fake news dataset with augmented fact-checking by chatgpt. *arXiv preprint arXiv:2312.11870*.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4461–4472.

Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023. Zero-shot rumor detection with propagation structure via prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5213–5221.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9781–9796. Association for Computational Linguistics.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of llms in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305.

Ruilin Luo, Tianle Gu, Haoling Li, Junzhe Li, Zicheng Lin, Jiayi Li, and Yujiu Yang. 2024. Chain of history: Learning and forecasting with llms for temporal knowledge graph completion. *arXiv preprint arXiv:2401.06072*.

Jiachen Ma, Yong Liu, Meng Han, Chunqiang Hu, and Zhaojie Ju. 2023a. Propagation structure fusion

for rumor detection based on node-level contrastive learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2023b. Kapalm: Knowledge graph enhanced language models for fake news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3999–4009.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles (version semeval-2020). https://doi.org/10.5281/zenodo.3952415. Accessed on YYYY-MM-DD.

Nikhil Mehta, María Leonor Pacheco, and Dan Goldwasser. 2022. Tackling fake news detection by continually improving social context representations using graph neural networks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1363–1380.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263.

Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. 2023. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15817–15835. Association for Computational Linguistics.

Nicholas Micallef, Bing He, Srijan Kumar, Mustaque Ahamad, and Nasir Memon. 2020. The role of the crowd in countering misinformation: A case study of the covid-19 infodemic. In *2020 IEEE international Conference on big data (big data)*, pages 748–757. IEEE.

Erxue Min and Sophia Ananiadou. 2023. Pesto: a post-user fusion network for rumour detection on social media. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 1–10.

Nishanth Nakshatri, Siyi Liu, Sihao Chen, Dan Roth, Dan Goldwasser, and Daniel Hopkins. 2023. Using llm for improving key event discovery: Temporal-guided news stream clustering with event summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4162–4173.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023a. Large language models and knowledge graphs: Opportunities and challenges. *TGDK*, 1(1):2:1–2:38.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023b. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1389–1403. Association for Computational Linguistics.

Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The surprising performance of simple baselines for misinformation detection. In *Proceedings of the Web Conference 2021*, pages 3432–3441.

Kellin Pelrine, Anne Imouza, Camille Thibault, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, Jean-François Godbout, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6399–6429. Association for Computational Linguistics.

Huyen Trang Phan, Ngoc Thanh Nguyen, and Dosam Hwang. 2023. Fake news detection: A survey of graph neural network methods. *Applied Soft Computing*, page 110235.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492.

Sudhakar Sengan, Subramaniyaswamy Vairavasundaram, Logesh Ravi, Ahmad Qasim Mohammad Al-Hamad, Hamzah Ali Alkhazaleh, and Meshal Alharbi. 2023. Fake news detection using stance extracted multimodal fusion-based hybrid neural network. *IEEE Transactions on Computational Social Systems*.

Omar Shaikh, Hongxin Zhang, William Held, Michael S. Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4454–4470. Association for Computational Linguistics.

Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. 2022. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Iftekharul Islam Shovon and Seokjoo Shin. 2023. The performance of graph neural network in detecting fake news from social media feeds. In *2023 International Conference on Information Networking (ICOIN)*, pages 560–564. IEEE.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019a. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 395–405.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big Data*, 8(3):171–188.

Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019b. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 436–439.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Harald Stiff and Fredrik Johansson. 2022. Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, 13(4):363–383.

Jianlin Su, Mingren Zhu, Ahmed Murtadha, Shengfeng Pan, Bo Wen, and Yunfeng Liu. 2022. Zlpr: A novel loss for multi-label classification. *arXiv preprint arXiv:2208.02955*.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2023a. Adapting fake news detection to the era of large language models. *arXiv preprint arXiv:2311.04917*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023b. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From chaos to clarity: Claim normalization to empower fact-checking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6594–6609.

Yoo Yeon Sung, Jordan L. Boyd-Graber, and Naeemul Hassan. 2023. Not all fake news is written: A dataset and analysis of misleading video headlines. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16241–16258. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Lin Tian, Xiuzhen Jenny Zhang, and Jey Han Lau. 2022. Duck: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adaku Uchendu, Jooyoung Lee, Hua Shen, Thai Le, Dongwon Lee, et al. 2023. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 11, pages 163–174.

Haoran Wang, Yingtong Dou, Canyu Chen, Lichao Sun, Philip S Yu, and Kai Shu. 2023. Attacking fake news detectors via manipulating news social engagement. In *Proceedings of the ACM Web Conference 2023*, pages 3978–3986.

Jiaying Wu and Bryan Hooi. 2023a. Decor: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2582–2593.

Jiaying Wu and Bryan Hooi. 2023b. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830*.

Xueqing Wu, Kung Hsiang Huang, Yi R Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 543–558. Association for Computational Linguistics (ACL).

Fan Xu, Pinyun Fu, Qi Huang, Bowei Zou, AiTi Aw, and Mingwen Wang. 2023a. Leveraging contrastive learning and knowledge distillation for incomplete modality rumor detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13492–13503.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023b. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 2501–2510.

Chang Yang, Peng Zhang, Wenbo Qiao, Hui Gao, and Jiaming Zhao. 2023a. Rumor detection on social media with crowd intelligence and chatgpt-assisted networks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5705–5717.

Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Zhiwei Yang. 2023b. WSDMS: debunk fake news via weakly supervised detection of misinforming sentences with contextualized social wisdom. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1525–1538. Association for Computational Linguistics.

Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2021. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 1417–1423.

Zhenrui Yue, Huimin Zeng, Yang Zhang, Lanyu Shang, and Dong Wang. 2023. Metaadapt: Domain adaptive few-shot misinformation detection via meta learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5223–5239. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

Fengzhu Zeng and Wei Gao. 2022. Early rumor detection using neural hawkes process with a new benchmark dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4105–4117.

Guixian Zhang, Shichao Zhang, and Guan Yuan. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, and Jianshu Chen. 2024. Thrust: Adaptively propels large language models with external knowledge. *Advances in Neural Information Processing Systems*, 36.

Xinyi Zhou, Kai Shu, Vir V Phoha, Huan Liu, and Reza Zafarani. 2022. "this is fake! shared it by mistake": Assessing the intent of fake news spreaders. In *Proceedings of the ACM Web Conference 2022*, pages 3685–3694.

You are a social media user. You are female. You are 18 to 29 years old. Racially, you are Hispanic. Financially, your annual family income is 30,000 to 74,999. Educationally, you are a college grad. Politically, you are a Republican. Meanwhile, you are probably registered to vote.

Table 4: An Example of a synthetic social media user prompt.

A Methodology Details

A.1 User Attribute Details

We simulate each synthetic user as an intersection of seven categories, and the detailed attribute descriptions of each category are as follows:

- **gender**: "*You are male.*"; "*You are female.*"

- **age**: "*You are under 17 years old.*"; "*You are 18 to 29 years old.*"; "*You are 30 to 49 years old.*"; "*You are 50 to 64 years old.*"; "*You are over 65 years old.*"

- **ethnicity**: "*Racially, you are White.*"; "*Racially, you are Black.*"; "*Racially, you are Hispanic.*"

- **education level**: "*Educationally, you are a college grad.*"; "*Educationally, you haven't graduated from college.*"; "*Educationally, you have a high school diploma or less.*"

- **family income**: "*Financially, your annual family income is more than 75,000.*"; "*Financially, your annual family income is 30,000 to 74,999.*"; "*Financially, your annual family income is less than 30,000.*"

- **political leaning**: "*Politically, you are a Republican.*"; "*Politically, you are a Democrat.*"

- **voter registration**: "*Meanwhile, you are registered to vote.*"; "*Meanwhile, you are probably registered to vote.*"; "*Meanwhile, you are not registered to vote.*"

We uniformly sample each value for the seven attributes to represent a social media user. We then concatenate these attributes prefixed with "*You are a social media user.*" as the prompt for the synthetic user. Table 4 illustrates a complete example of a synthetic social media user prompt.

Algorithm 1: Pseudo-code of user-news network generation.

Data: news content s; graph size m; α to control the probability of commenting on the news; β to control the balance of tree height and width; k to control candidate set size

Result: user-news network $\mathcal{G}(\mathcal{V}, \mathcal{E})$

1   $\mathcal{V} = [s]$
2   $\mathcal{E} = []$
3   $\mathcal{F} = [s]$
4   $\mathcal{H}$ denoting height of each node
5   $\mathcal{W}$ denoting width of each node
6   **while** $\|V\| \leq m$ **do**
7     $\boldsymbol{u} = \psi()$
8     $p \sim U(0, 1)$
9   **if** $p <= \alpha$ **then**
10       $\text{PROMPT} = \phi_1(\boldsymbol{s}, \boldsymbol{u})$
11       $\boldsymbol{f} = \boldsymbol{s}$
12   **else**
13       $\mathcal{P} = \beta * \mathcal{H} + (1 - \beta) * \mathcal{W}$
14       $\{\boldsymbol{c}^i\}_i = \text{Sample}(\mathcal{V}, \mathcal{P}, k)$
15       $\text{SELECT} = \phi_3(\boldsymbol{s}, \boldsymbol{u}, \{\boldsymbol{C}^i\}_i)$
16       $j = \text{LLM}(\text{SELECT})$
17       $\boldsymbol{f} = \boldsymbol{C} = \boldsymbol{C}^j$
18       $\text{PROMPT} = \phi_2(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{C})$
19   **end**
20     $\boldsymbol{s}_{out} = \text{LLM}(\text{PROMPT})$
21     $\mathcal{V}.\text{append}(\boldsymbol{s}_{out})$
22     $\mathcal{E}.\text{append}((\boldsymbol{s}_{out}, \boldsymbol{f}))$
23     $\mathcal{F}.\text{append}(\boldsymbol{f})$
24   update $\mathcal{H}$ and $\mathcal{W}$
25   **end**
    Return: $\mathcal{G}(\mathcal{V}, \mathcal{E})$

A.2   User-News Networks Details

Our generated user-news interaction networks $\mathcal{G}$ forms a tree structure, where $\|\mathcal{V}\| = \|\mathcal{E}\| + 1$. To control the iterative process of generating user-news networks, we design hyperparameters α to control the probability of commenting on news and β to control the balance of tree height and width. We present an algorithmic summary of the generation process in Algorithm 1, where $\phi_1(\boldsymbol{s}, \boldsymbol{u})$, $\phi_2(\boldsymbol{s}, \boldsymbol{u}, \boldsymbol{C})$, and $\phi_3(\boldsymbol{s}, \boldsymbol{u}, \{\boldsymbol{C}^i\}_i)$ denotes *comment on news*, *Comment on a comment*, and *Select a comment to comment* prompt generation process, $\psi()$ denotes the diverse user attribute prompt generation process as mentioned in Appendix A.1, $\text{Sample}(\mathcal{V}, \mathcal{P}, k)$ denote the sample function that

sample k instances from $\mathcal{V}$ according to the probability $\mathcal{P}$. Table 9 presents the prompt templates.

A.3   Explainable Proxy Task Details

We propose four proxy tasks to enhance news articles, the details of **Sentiment Analysis**, **Framing Detection**, and **Propaganda Tactics Detection** are as follows:

- **Sentiment Analysis**: we employ six basic emotions: *anger*; *disgust*; *fear*; *happiness*; *sadness*; and *surprise*.

- **Framing Detection**: we employ 14 news frames: *Economic*; *Capacity and resources*; *Morality*; *Fairness and equality*; *Legality, constitutionality and jurisprudence*; *Policy prescription and evaluation*; *Crime and punishment*; *Security and defense*; *Health and safety*; *Quality of life*; *Cultural identity*; *Among public opinion*; *Political*; *External regulation and reputation*.

- **Propaganda Tactics Detection**: we employ 19 propaganda tactics: *Conversation Killer*; *Whataboutism*; *Doubt*; *Straw Man*; *Red Herring*; *Loaded Language*; *Appeal to Fear-Prejudice*; *Guilt by Association*; *Flag Waving*; *False Dilemma-No Choice*; *Repetition*; *Appeal to Popularity*; *Appeal to Authority*; *Name Calling-Labeling*; *Slogans*; *Appeal to Hypocrisy*; *Exaggeration-Minimisation*; *Obfuscation-Vagueness-Confusion*; *Causal Oversimplification*.

Table 10 presents the presents the prompt template of each proxy task.

A.4   LLM-Based Expert Ensemble Details

We propose three LLM-based approaches to selectively integrate the prediction of each expert. The description $\boldsymbol{d}_i$ of each expert e_i is as follows:

- **w/o expert**: *This expert is comprehensive.*

- **sentiment**: *This expert focuses on the emotion of this news.*

- **framing**: *This expert focuses on the framing of this news.*

- **propaganda tactics**: *This expert focuses on the propaganda tactics of this news..*

- **retrieval**: *This expert focuses on the external knowledge of this news.*

- **stance**: *This expert focuses on the stance of related comments on this news.*

- **relation**: *This expert focuses on the relation of related comments on this news.*

to obtain the confidence score, we employ a softmax operator *score* for binary classification and absolute value operator for multi-label classification. We provide the prompt templates in Table 11.

B   Experiment Setting Details

B.1   Dataset Details

We evaluate DELL and baselines on three tasks related to fake news detecton.
   1) Fake News Detection:

- **Pheme** (Buntain and Golbeck, 2017) is a dataset of potential rumors on Twitter and journalistic assessments of their accuracies.

- **LLM-mis** (Chen and Shu, 2023) is a LLM-generated misinformation dataset with different LLM generators and generation approaches.

2) Framing Detection:

- **MFC** (Card et al., 2015b) contains labeled and unlabeled articles on six issues from 14 newspapers covering the years 1990-2014, though some issues have broader coverage. The issues include climate; the death penalty; gun control; immigration; same-sex sex; and tobacco. We sample the labeled articles as a benchmark.

- **SemEval-23F** (Piskorski et al., 2023) aim to identify one or more frames used in an article from a pool of 14 generic frames: Security and defense; Fairness and equality; Political; Capacity and resources; Economic; Morality; Policy prescription and evaluation; Legality Constitutionality and jurisprudence; External regulation and reputation; Quality of life; Health and safety; Cultural identity; Crime and punishment; and Public opinion.

3) Propaganda Tactic Detection:

- **Generated** is a benchmark generated by *ChatGPT*. We first determine 4 topics: Russia and Ukraine; Palestine and Israel; the Republican Party; and the Democratic Party. Around these topics, we generated 5 paragraphs for each tactic.

- **SemEval-20** (Martino et al., 2020) contains 14 possible propaganda tactics: Appeal to fear-prejudice; Black-and-White Fallacy; Name Calling, Labeling; Slogans; Whataboutism,Straw Men,Red Herring; Exaggeration, Minimisation; Loaded Language; Repetition; Causal Oversimplification; Bandwagon,Reductio ad hitlerum; Flag-Waving; Thought-terminating Cliches; Appeal to Authority; and Doubt. This benchmark merges some tactics into one category.

- **SemEval-23F** (Piskorski et al., 2023) contains 6 main categories: Attack on reputation; Justification; Simplification; Distraction; Call; and Manipulative wording. It contains 19 propaganda tactics: Conversation Killer; False Dilemma-No Choice; Appeal to Popularity; Doubt; Flag Waving; Slogans; Whataboutism; Straw Man; Loaded Language; Name Calling-Labeling; Obfuscation-Vagueness-Confusion; Appeal to Fear-Prejudice; Causal Oversimplification; Red Herring; Repetition; Exaggeration-Minimisation; Appeal to Authority; Guilt by Association; and Appeal to Hypocrisy.

   We randomly sample 1,000 instances from each benchmark (select all if there are less than 1,000 instances), and divided the training set, validation set, and test set according to the ratio of 7:2:1.

B.2   Bseline Details

- ZERO-SHOT asks LLMs to conduct detection.

- FEW-SHOT first provides LLMs with some pairs of news instances and labels and then asks LLMs to conduct detection.

- RETRIEVAL-augmented generation first provides LLMs with the external knowledge retrieved from Wikipedia, which is the same as **Knowledge Retrieval** proxy task. It then asks LLM to conduct detection.

- F3 *Z-CoT* (Lucas et al., 2023) uniquely leverages LLMs' self-formulated rationales by integrating a standard instruction with the simple phrase, "Let's think step by step known as Chain of Thoughts (CoT)."

- F3 *DeF-Gen* (Lucas et al., 2023) focuses contextually, emphasizing deductive and abductive reasoning.

- TAPE W/O GRAPH (He et al., 2023b) focus on leveraging LLMs to capture textual information as features, which can subsequently enhance GNN performance on downstream tasks. Here we only employ the textual information generated by LLMs to enhance news content. DEBERTA (He et al., 2021) leverages the pretrained language model DeBERTa to encode news content, then feed them into an MLP classifier.

- K-HOPS (Huang et al., 2023a) incorporates randomly selected neighbors into the prompt, where the idea behind is to aggregate information from neighboring nodes, following GCN.

- K-ATTENTION (Huang et al., 2023a) is designed to weigh the influence of neighboring nodes during the prediction process, following GAT.

- TAPE W/ GRAPH (He et al., 2023b) puts the enhanced news content into the user-news network and employs graph neural networks to conduct detection.

- GCN (Kipf and Welling, 2017) adopt multiple GNN layers and a meaning pooling to obtain the user-news network representations.

- DEFEND (Shu et al., 2019a) conducts explainable detection by the attention weights, we set maximum sentence length and maximum comment length as 96, maximum sentence count as 64, and maximum comment count as 10 to reproduce so that the approach is applicable to our tasks and datasets.

- RVNN (Ma et al., 2018) proposes two recursive neural model stratages: bottom-up and top-down tree-structured neural networks. We employ the top-down structure.

- HYPEHN (Grover et al., 2022) is a discourse-aware hyperbolic spectral co-attention network. It is a fusion of hyperbolic graph representation learning with a novel Fourier co-attention mechanism in an attempt.

- GET (Xu et al., 2022) models claims and related evidences as graph-structured data and capture the long-distance semantic dependency among dispersed relevant snippets via neighborhood propagation.

Hyperparameter	Value
optimizer	Adam
learning rate	10^{-4}
weight decay	10^{-5}
dropout	0.5
hidden dim	1024
embedding dim	1024
GNN layers	2
maximum epochs	100
temperature for reaction generation τ	0.6
temperature	0.1

Table 5: Hyperparameter settings of DELL.

- WSDMS (Yang et al., 2023b) needs bag-level labels for training but possesses the capability to infer both sentence-level misinformation and article-level veracity, facilitated by pertinent social media conversations meticulously contextualized with news sentences.

For the LLM-based baseline, we provide the prompt templates in Table 12. Each baseline prompt template contains a task-related prompt to describe the task and a baseline-related prompt.

B.3 Hyperparameters

The hyperparameter settings of DELL are presented in Table 5 to facilitate reproduction.

C Additional Results

We leverage *Mistral-7B* (Jiang et al., 2023a), *LLaMA2-70B* (Touvron et al., 2023), and *ChatGPT* as the base LLMs. Tabel 7 presents more results. DELL still outperforms other baselines.

For context, DELL has three components: Diverse Reaction Generation; Explainable Proxy Tasks; and LLM-Based Expert Ensemble. On the component level, we conduct more ablation studies as follows:

- w/o Diverse Reaction Generation: we remove the network part and only employ the news content and related proxy task outputs.

- w/o Explainable Proxy Tasks: we remove all proxy tasks and conduct experiments with news content and generated content.

- w/o LLM-Based Expert Ensemble: we replace LLM-based ensembling with simple majority

voting: majority vote; confidence weighted sum; and train the weights on validation set.

We present the results of the ablation study in Table 6. Every module of DELL could improve the fake news detection performance.

D   Additional Analysis

D.1   Model Robustness to Comments (cont.)

Figure 9 presents the results of the other six benchmarks. On other benchmarks, DELL drops the least in performance with reduced comments. Specifically, DELL almost keeps the same on **Pheme** and drops 11.8% on **MFC**, 1.1% on **SemEval-23F**, 7.1% on **Generated**, 3.9% on **SemEval-20**, and 3.3% on **SemEval-23P**.

D.2   Expert Selection (cont.)

Figure 10 presents the results of the other six benchmarks. Besides this experiment, we also examine the count of experts in every selection and corresponding performance. The results are shown in Table 8.

D.3   Case Study (cont.)

Table 14, 15, and 16 provide more cases of the explanations of proxy tasks generated by DELL. It illustrates that LLMs could generate reasonable explanations of proxy tasks, providing more information in identifying fake news.

Method	Fake News Detection				Framing Detection				Propaganda Tactic Detection					
	Pheme		LLM-mis		MFC		SemEval-23F		Generated		SemEval-20		SemEval-23P	
	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF
DELL	.820	.820	.928	.930	.509	.603	.572	.718	.598	.577	.525	.636	.386	.643
w/o Diverse Reaction *Single*	.790	.790	.896	.900	.433	.575	.528	.663	.551	.552	.516	.602	.407	.604
w/o Diverse Reaction *Vanilla*	.800	.800	.907	.910	.440	.581	.521	.673	.522	.519	.490	.599	.370	.619
w/o Diverse Reaction *Confidence*	.789	.790	.875	.880	.429	.581	.361	.544	.566	.588	.524	.613	.376	.634
w/o Diverse Reaction *Selective*	.810	.810	.887	.890	.477	.594	.521	.670	.528	.537	.542	.629	.365	.606
w/o Explainable Proxy Tasks	.790	.790	.915	.920	.417	.577	.518	.704	.543	.556	.504	.596	.364	.620
Majority Vote	.830	.830	.917	.920	.418	.576	.555	.703	.580	.593	.544	.647	.377	.644
Confidence weight	.820	.820	.917	.920	.458	.593	.583	.705	.578	.550	.504	.613	.369	.661
Train on Validation Set	.800	.800	.897	.900	.496	.585	.579	.674	.566	.527	.546	.622	.407	.638

Table 6: Performance of variants of DELL. The ablation study results illustrate that every module of DELL is helpful for fake news detection.

Method	Fake News Detection				Framing Detection				Propaganda Tactic Detection					
	Pheme		LLM-mis		MFC		SemEval-23F		Generated		SemEval-20		SemEval-23P	
	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF	MaF	MiF
LLM-based baselines with *Mistral-7B*.														
ZERO-SHOT	.450	.450	.517	.560	.247	.265	.371	.431	.189	.202	.293	.408	.207	.274
FEW-SHOT	.385	.390	.639	.660	.259	.303	.376	.432	.170	.217	.382	.466	.306	.372
RETRIEVAL	.466	.480	.648	.670	.216	.260	.285	.383	.116	.134	.280	.389	.234	.310
TAPE W/O GRAPH	.789	.790	.895	.900	.278	.497	.409	.610	.340	.353	.320	.595	.259	.632
K-HOPS	.301	.430	.533	.550	.255	.273	.377	.489	.107	.110	.286	.406	.156	.243
K-ATTENTION	.310	.420	.483	.510	.280	.336	.414	.508	.143	.145	.433	.474	.264	.312
TAPE W GRAPH	.800	.800	.897	.900	.270	.485	.379	.633	.341	.358	.328	.598	.253	.608
LLM-based baselines with *LLaMA2-70B*.														
ZERO-SHOT	.403	.410	.650	.650	.331	.374	.380	.493	.176	.178	.143	.228	.044	.140
FEW-SHOT	.322	.420	.670	.670	.312	.357	.396	.480	.117	.119	.404	.452	.335	.415
RETRIEVAL	.513	.520	.672	.680	.315	.354	.325	.483	.163	.167	.109	.186	.073	.150
TAPE W/O GRAPH	.748	.750	.856	.860	.376	.581	.448	.654	.381	.427	.358	.613	.244	.612
K-HOPS	.310	.420	.634	.690	.327	.386	.394	.521	.204	.180	.189	.292	.054	.167
K-ATTENTION	.400	.410	.750	.760	.306	.378	.414	.539	.198	.196	.312	.429	.109	.203
TAPE W GRAPH	.752	.760	.857	.860	.392	.575	.453	.670	.370	.420	.320	.592	.244	.632
All expert performance of DELL with *ChatGPT*.														
VANILLA	.790	.790	.915	.920	.417	.577	.518	.704	.543	.556	.504	.596	.364	.620
SENTIMENT	.780	.780	.867	.870	.413	.552	.536	.684	.510	.492	.499	.578	.343	.650
FRAMING	.810	.810	.887	.890	.446	.571	.509	.658	.509	.541	.520	.613	.375	.618
PROPAGANDA	.780	.780	.858	.860	.458	.598	.487	.604	.506	.525	.496	.583	.363	.606
RETRIEVAL	.779	.780	.897	.900	.450	.570	.512	.646	.522	.520	.513	.589	.370	.601
STANCE	.780	.780	.917	.920	.435	.571	.532	.683	.517	.547	.496	.606	.376	.631
RESPONSE	.780	.780	.928	.930	.428	.582	.506	.695	.536	.538	.493	.618	.364	.646
expert ensemble of DELL with *Mistral-7B*.														
DELL *Vanilla*	.770	.770	.888	.890	.411	.526	.577	.689	.551	.519	.513	.618	.337	.566
DELL *Confidence*	.789	.790	.866	.870	.458	.571	.539	.676	.539	.507	.484	.590	.347	.623
DELL *Selective*	.820	.820	.917	.920	.478	.579	.570	.700	.608	.577	.493	.608	.367	.662
expert ensemble of DELL with *LLaMA2-70B*.														
DELL *Vanilla*	.722	.730	.906	.910	.453	.582	.549	.700	.579	.553	.563	.655	.382	.646
DELL *Confidence*	.624	.670	.894	.900	.421	.569	.509	.685	.555	.549	.541	.649	.371	.632
DELL *Selective*	.810	.810	.897	.900	.457	.592	.573	.704	.575	.547	.505	.615	.367	.655

Table 7: Performance of DELL and baselines using other LLMs on seven datasets from three tasks related to fake news detection. DELL still outperforms other baselines.

Benchmark		0	1	2	3	4	5	6	7
Pheme	instance count	0	1	5	49	22	19	4	0
	macro f1-score	nan	1.00	1.00	.816	.818	.789	.750	nan
LLM-mis	instance count	0	0	2	37	32	16	9	4
	macro f1-score	nan	nan	.500	.946	.813	1.00	.889	1.00
MFC	instance count	0	0	3	43	36	17	0	1
	macro f1-score	nan	nan	.571	.602	.548	.583	nan	.800
SemEval-23F	instance count	0	0	2	11	15	12	8	4
	macro f1-score	nan	nan	63.2	62.1	70.5	73.3	66.7	68.8
Generated	instance count	0	0	1	27	23	19	4	0
	macro f1-score	nan	nan	1.00	.704	.500	.476	.600	nan
SemEval-20	instance count	0	1	2	8	15	8	4	0
	macro f1-score	nan	0.833	0.800	0.615	0.604	0.615	0.667	nan
SemEval-23P	instance count	0	1	2	8	19	13	7	4
	macro f1-score	nan	1.00	.625	.582	.682	.607	.644	.745

Table 8: The count of experts in every selection and corresponding performance in the *selective* approach.

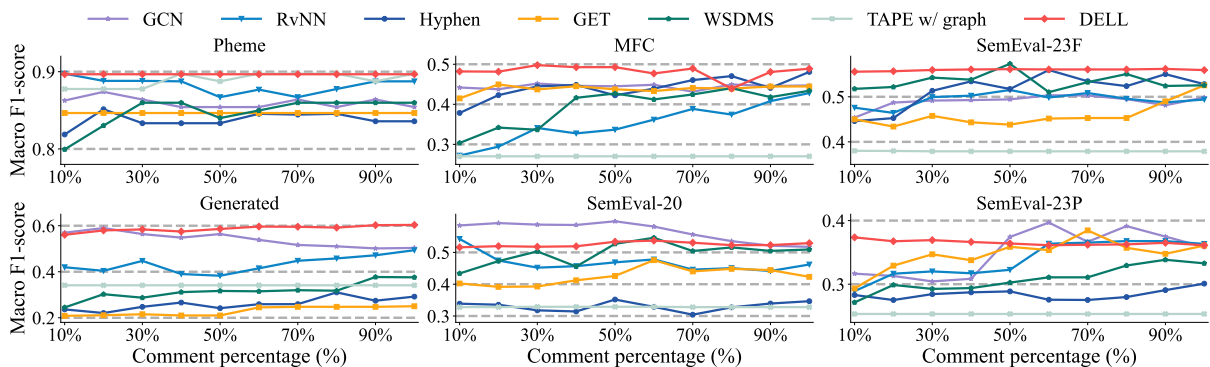
Figure 9: Performance of DELL and baselines when the comments are gradually removed. DELL shows great robustness to the availability of comments.

Comment on news	u You view a piece of news with the following content. News: s Task: Please comment on this news on social media. Your comment is limited to 40 words. Your comment:
Comment on a comment	u You view a piece of news and a related comment chain on social media, and their contents are as follows. News: s Comment 1: c_1 Comment 2: c_2 $\ldots$ Comment k: c_k Task: Please reply to the last comment(comment k) on social media. Your reply is limited to 40 words. Your reply:
Select a comment to comment	u You view a piece of news and related comment chains on social media, and their contents are as follows. News: s Comment Chain 1: Comment 1: c_1^1 Comment 2: c_2^1 $\ldots$ Comment k_1: $c_{k_1}^1$ Comment Chain 2: $\ldots$ Comment Chain n: $\ldots$ Comment k_n: $c_{k_n}^n$ Task: Please select a comment chain that you would most like to comment on. Answer the selected number and explain the reason. Answer:

Table 9: Prompt templates of generating user-news networks

Sentiment Analysis	News: s Task: Which emotions does the news contain? Please choose the three most likely ones: anger, disgust, fear, happiness, sadness, and surprise. Please provide your reasoning. Answer:
Framing Detection	News: s Task: Framing is a strategic device and a central concept in political communication for representing different salient aspects and perspectives to convey the latent meaning of an issue. Which framings does the news contain? Please choose the five most likely ones: Economic; Capacity and resources; Morality; Fairness and equality; Legality, constitutionality and jurisprudence; Policy prescription and evaluation; Crime and punishment; Security and defense; Health and safety; Quality of life; Cultural identity; Among public opinion; Political; External regulation and reputation. Please provide your reasoning. Answer:
Propaganda Tactics Detection	News: s Task: Propaganda Tactics are methods used in propaganda to convince an audience to believe what the propagandist wants them to believe. Which propaganda techniques does the news contain? Please choose the five most likely ones: Conversation Killer; Whataboutism; Doubt; Straw Man; Red Herring; Loaded Language; Appeal to Fear-Prejudice; Guilt by Association; Flag Waving; False Dilemma-No Choice; Repetition; Appeal to Popularity; Appeal to Authority; Name Calling-Labeling; Slogans; Appeal to Hypocrisy; Exaggeration-Minimisation; Obfuscation-Vagueness-Confusion; Causal Oversimplification. Please provide your reasoning. Answer:
Knowledge Retrieval	News: s Task: Identify five named entities within the news above that necessitate elucidation for the populace to understand the news comprehensively. Ensure a diverse selection of the entities. The answer should in the form of python list. Answer:
Stance Detection	Task: Determine the stance of sentence 2 on sentence 1. Is it supportive, neutral or opposed? Provide your reasoning. Sentence 1: s_1 Sentence 2: s_2 Answer:
Response Characterization	Sentence 1: s_1 Sentence 2: s_2 Task: Sentence 1 and Sentence 2 are two posts on social networks. Please judge whether the sentence 2 replies to the sentence 1. Answer yes or no and provide the reasoning. Answer:

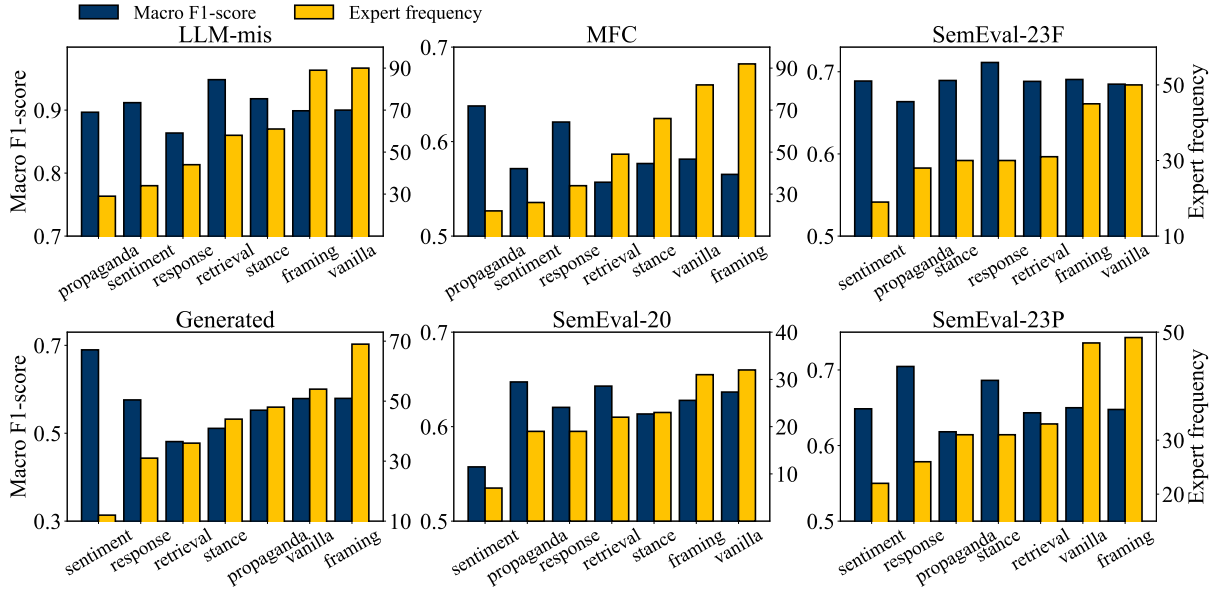Table 10: Prompt templates of each proxy task.

Figure 10: The frequency of expert selections and performance when a particular expert is selected in the *Selective* approach.

Vanilla	News: s Some experts give predictions about the news. Expert 1: d_1. The expert predicts the label of this news is ℓ_1. Expert 2: d_2. The expert predicts the label of this news is ℓ_2. . . . Expert 7: d_7. The expert predicts the label of this news is ℓ_7. Question: Based on the analysis of experts, please judge the final label of this news. Give the label in the form of "[your answer]", do not give any explanation. Label:
Confidence	News: s Some experts give predictions about the news. Expert 1: d_1. The expert predicts the label of this news is ℓ_1. The confidence scores are $score_1$. Expert 2: d_2. The expert predicts the label of this news is ℓ_2. The confidence scores are $score_2$. . . . Expert 7: d_7. The expert predicts the label of this news is ℓ_7. The confidence scores are $score_7$. Question: Based on the analysis of experts, please judge the final label of this news. Give the label in the form of "[your answer]", do not give any explanation. Label:
Selective	News: s Expert 1: d_1. Expert 2: d_2. . . . Expert 7: d_7. To understand this news, which expert knowledge do you need? Return a Python list, e.g. [expert 1, expert 2, expert 6].

Table 11: Prompt templates of our proposed LLM-based ensemble approaches.

Fake News Detection	Task: Please determine whether the news is real or fake.
Framing Detection	Task: Framing is a strategic device and a central concept in political communication for representing different salient aspects and perspectives to convey the latent meaning of an issue. Which framings does the news contain? Please choose from: *the candidate label set of a specific dataset*.
Propaganda Tactic Detection	Task: Propaganda techniques are methods used in propaganda to convince an audience to believe what the propagandist wants them to believe. Which propaganda techniques does the news contain? Please choose from: *the candidate label set of a specific dataset*.
ZERO-SHOT	*Task-related prompt* News: s Answer:
FEW-SHOT	*Example pairs of news and lable* *Task-related prompt* News: s Answer:
RETRIEVAL	Knowledge: *externel knowledge retrieved from Wikipedia* *Task-related prompt* News: s Answer:
TAPE	News: s *Task-related prompt* Provide your reasoning. Answer:
K-HOPS	News: s It has the following comments: *the comments related to the news* *Task-related prompt* Answer:
K-ATTENTION	News: s It has the following comments: *the comments related to the news* Task: Please return the index of the comment that best assists in understanding this news. Answer: News: s It has the following comments: *the selected comments* *Task-related prompt* Answer:

Table 12: Prompt templates of LLM-based baselines.

F3 *Z-CoT*	You are an AI assistant trained to detect fake news. Deeply Analyze the given text, think step-by-step, and determine if it is real or fake news. {News content}
	You are an AI assistant trained to detect framing. Deeply Analyze the given text, think step-by-step, which framings does the following news contain? Please choose from: {label set} {News content}
	You are an AI assistant trained to detect propaganda tactic. Deeply Analyze the given text, think step-by-step, which propaganda tactics does the following news contain? Please choose from: {label set} {News content}
F3 *DeF-Gen*	You are an AI assistant trained to detect fake news. Determine if the text contains factual information supported by evidence (real) or 'misleading/inaccurate information (fake). Respond with 'True' or 'False'. {News content}
	You are an AI assistant trained to detect framing. Framing is a strategic device and a central concept in political communication for representing different salient aspects and perspectives to convey the latent meaning of an issue. Which framings does the following news contain? Please choose from:{label set} {News content}
	You are an AI assistant trained to detect propaganda tactic. Propaganda tactics are methods used in propaganda to convince an audience to believe what the propagandist wants them to believe. Which propaganda tactics does the following news contain? Please choose from: {label set} {News content}

Table 13: Prompt templates of LLM-based baselines. (cont.)

	News:
	Given the recent wave of allegations pertaining to sexual misconduct on the part of congressmen in Washington, it is not inconceivable that similar accusations could be directed at Senator Chuck Schumer. While I acknowledge that some of my previous posts may have been hasty and not grounded in factual evidence, I must clarify that I do not make such remarks maliciously. I hope this insight assuages any potential doubts about my intentions.

News:

Given the recent wave of allegations pertaining to sexual misconduct on the part of congressmen in Washington, it is not inconceivable that similar accusations could be directed at Senator Chuck Schumer. While I acknowledge that some of my previous posts may have been hasty and not grounded in factual evidence, I must clarify that I do not make such remarks maliciously. I hope this insight assuages any potential doubts about my intentions.

User profile:

You are a social media user. You are female. You are 18 to 29 years old. Racially, you are White. Financially, your annual family income is less than 30,000. Educationally, you haven't graduated from college. Politically, you are a Democrat. Meanwhile, you are not registered to vote.

Comment:

As a fellow Democrat, I urge Senator Schumer to address these allegations swiftly and transparently. It's important that we hold all elected officials accountable for their actions.

Sentiment Analysis	Based on the content of the news, the three most likely emotions are anger, disgust, and fear. Anger is likely to be felt by those who believe that Senator Schumer has engaged in improper behavior and that he should be held accountable. Disgust is also likely to be felt by those who are repulsed by the idea of sexual misconduct on the part of any politician, regardless of their political affiliation. Fear is likely to be felt by those who are concerned about the potential repercussions of these allegations for the broader political landscape, as well as for the safety and security of the senator himself.
Framing Detection	The news contains the following five likely framings: 1. Morality: The allegations of sexual misconduct against congressmen in Washington are often framed as a moral issue, with some commentators arguing that such behavior is unethical and unacceptable. 2. Fairness and equality: The issue of sexual misconduct is also framed as a matter of fairness and equality, with some arguing that women should be treated with respect and dignity, regardless of their position or status. 3. Legality, constitutionality and jurisprudence: The legal implications of sexual misconduct are also a key concern, with some commentators arguing that such behavior is a violation of federal law and the Constitution. 4. Policy prescription and evaluation: The issue of sexual misconduct is also framed as a matter of policy prescription and evaluation, with some commentators arguing that Congress should take action to address the problem and prevent it from happening in the future. 5. External regulation and reputation: The issue of sexual misconduct is also framed as a matter of external regulation and reputation, with some commentators arguing that Congress should be held accountable for its actions and that the public should demand better behavior from its elected representatives.

Table 14: Examples of the explanations of proxy tasks generated by DELL.

Propaganda Tactics Detection	The news contains the following five propaganda techniques: 1. Appeal to Authority: The news uses the authority of the senator to make the audience believe that he is trustworthy and reliable. 2. Appeal to Popularity: The news uses the popularity of the senator to make the audience believe that he is a good person and deserves their support. 3. Appeal to Fear-Prejudice: The news uses the fear of being accused of sexual misconduct to make the audience believe that the senator is innocent and deserves their support. 4. False Dilemma-No Choice: The news presents the audience with a false dilemma by suggesting that the only choice is to support the senator or to believe the allegations against him. 5. Exaggeration-Minimisation: The news exaggerates the importance of the allegations against the senator and minimizes the evidence supporting them.
Stance Detection	The stance of sentence 2 on sentence 1 is supportive. The speaker of sentence 2, who is a fellow Democrat, acknowledges the potential allegations against Senator Schumer and urges him to address them swiftly and transparently. This shows that the speaker supports the idea of holding elected officials accountable for their actions, which is a sentiment that is expressed in sentence 1. Additionally, the speaker's use of the word fellowsuggests that they share a common identity with Senator Schumer, further indicating support.
Response Characterization	Yes, sentence 2 replies to sentence 1. Sentence 1 expresses the possibility that Senator Chuck Schumer could be accused of sexual misconduct, while sentence 2 urges him to address the allegations in a transparent manner. Sentence 1 and Sentence 2 are related in that they both discuss Senator Schumer and the allegations against him.

Table 15: Examples of the explanations of proxy tasks generated by DELL. (cont.)

| Knowledge Retrieval | Given the recent wave of allegations pertaining to sexual misconduct (Sexual misconduct is misconduct of a sexual nature which exists on a spectrum that may include a broad range of sexual behaviors considered unwelcome. This includes conduct considered inappropriate on an individual or societal basis of morality, sexual harassment and/or criminal sexual assault. However generally, from a purely legal standpoint, sexual misconduct is a ÏayÌ termẅhich represents a boundary that has been broken, dictated by a moral set of conduct, particularly where the situation is normally non-sexual and therefore unusual for sexual behavior, or where there is some aspect of personal power or authority that makes sexual behavior inappropriate.) on the part of congressmen (A member of congress (MOC) is a person who has been appointed or elected and inducted into an official body called a congress, typically to represent a particular constituency in a legislature. The term member of parliament (MP) is an equivalent term within a parliamentary system of government.) in Washington (George Washington (February 22, 1732 Ž2013 December 14, 1799) was an American Founding Father, military officer, and statesman who served as the first president of the United States from 1789 to 1797. Appointed by the Second Continental Congress as commander of the Continental Army in June 1775, Washington led Patriot forces to victory in the American Revolutionary War and then served as president of the Constitutional Convention in 1787, which drafted and ratified the Constitution of the United States and established the American federal government. Washington has thus been called the Ëather of his CountryÏ.) it is not inconceivable that similar accusations could be directed at Senator (A senate is a deliberative assembly, often the upper house or chamber of a bicameral legislature. The name comes from the ancient Roman Senate (Latin: Senatus), so-called as an assembly of the senior (Latin: senex meaning Ïhe elderÖr Öld manÏ) and therefore considered wiser and more experienced members of the society or ruling class. However the Roman Senate was not the ancestor or predecessor of modern parliamentarism in any sense, because the Roman senate was not a de jure legislative body.Many countries have an assembly named a senate, composed of senators who may be elected, appointed, have inherited the title, or gained membership by other methods, depending on the country.) Chuck Schumer (Charles Ellis Schumer (SHOO-mŏ259r; born November 23, 1950) is an American politician serving as Senate Majority Leader since 2021 and the senior United States senator from New York since 1999. A member of the Democratic Party, he has led the Senate Democratic Caucus since 2017 and was Senate Minority Leader from 2017 to 2021. Schumer is in his fifth Senate term, making him the longest-serving US senator from New York, having surpassed Daniel Patrick Moynihan and Jacob K. Javits in 2023.). While I acknowledge that some of my previous posts may have been hasty and not grounded in factual evidence, I must clarify that I do not make such remarks maliciously. I hope this insight assuages any potential doubts about my intentions. |

Table 16: Examples of the explanations of proxy tasks generated by DELL. (cont.)