TAGFN: A Text-Attributed Graph Dataset for Fake News Detection in the Age of LLMs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have recently revolutionized machine learning on text-attributed graphs, but the application of LLMs to graph outlier detection, particularly in the context of fake news detection, remains significantly underexplored. One of the key challenges is the scarcity of large-scale, realistic, and well-annotated datasets that can serve as reliable benchmarks for outlier detection. To bridge this gap, we introduce TAGFN, *a large-scale, real-world text-attributed graph dataset for outlier detection*, specifically fake news detection. TAGFN enables rigorous evaluation of both traditional and LLM-based graph outlier detection methods. Furthermore, it facilitates the development of misinformation detection capabilities in LLMs through fine-tuning. We anticipate that TAGFN will be a valuable resource for the community, fostering progress in robust graph-based outlier detection and trustworthy AI. The dataset is available at `https://huggingface.co/datasets/anonymous-tagfn/TAGFN` and our code is available at `https://anonymous.4open.science/r/tagfn`.

1 Introduction

Graph-structured data offers a flexible framework for modeling interactions among entities in diverse domains such as social networks, recommender systems, and biological networks (Xiao et al., 2020; Liu et al., 2023; Li et al., 2025a). In many practical applications, nodes are often associated with rich textual attributes, forming text-attributed graphs (TAGs) (Yang et al., 2021; Yan et al., 2023). By integrating structural and semantic information, TAGs enable more fine-grained learning.

A growing body of work has explored the integration of graph learning with large language models (LLMs) for TAGs, yielding impressive results on tasks such as node classification (Chen et al., 2023; Zhu et al., 2025), out-of-distribution detection (Xu et al., 2025a,b), and question answering (Yasunaga et al., 2022; Li et al., 2025b). Among the diverse applications of outlier detection, misinformation detection emerges as a natural use case for TAGs. The semantical information in news and user content, combined with the propagation graph structure, collectively provides critical signals for identifying fake news. Despite the promise of LLMs for graph learning (Li et al., 2023; Chen et al., 2024), their application to outlier detection on graphs (Liu et al., 2022)–particularly for fake news detection and misinformation detection–remains largely underexplored. Most existing outlier detection methods are developed for graphs without textual information, leaving outlier detection on TAGs significantly understudied.

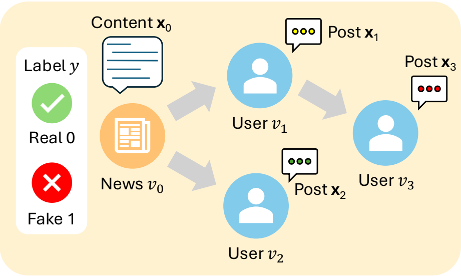

Figure 1: A toy example of news propagation graph in TAGFN, where the root node denotes the news and child nodes represent users, each attributed with text.

In this context, the synergy between the graph structure and textual attributes plays a critical role in understanding phenomena such as information dissemination and the emergence of outliers (e.g., fake news). However, a primary obstacle to investigating outlier detection on TAGs is *the scarcity of large-scale, realistic, and well-annotated datasets that combine graph structure with meaningful textual attributes and reliable ground-truth labels*. Existing benchmarks are limited in scale, lack raw textual attributes, or fail to provide realistic ground-truth labels necessary for the rigorous evaluation of outlier detection methods.

Table 1: Comparison of TAGFN with existing datasets.

	UPFD (2021)	CS-TAG (2023)	AD-LLM (2024)	NLP-ADBench (2024)	TAGFN
Text	✗	✓	✓	✓	✓
Outlier	✓	✗	✓	✓	✓
Large	✗	✓	✗	✗	✓
Graph	✓	✓	✗	✗	✓
Time	✗	✗	✗	✗	✓

To bridge this gap, we introduce TAGFN, a large-scale, real-world TAG dataset for outlier detection in the context of fake news detection. As shown in Figure 1, each graph in TAGFN depicts the propogation of a news, where the root node represents the news itself and child nodes represent the users who propagated it. Each node is attributed with text–either the news content or user posts–capturing the multifaceted nature of information disseminations. We also provide ground-truth outlier labels for each graph/news, indicating whether the news is fake or real. We anticipate that TAGFN will serve as a valuable resource for the research community, catalyzing progress at the intersection of LLMs, graph learning, and misinformation detection. By enabling systematic evaluation and development of advanced models, TAGFN aims to advance the state of the art in both graph machine learning and trustworthy AI.

Our contributions are mainly as follows:

- We construct TAGFN, the first large-scale, real-world TAG dataset tailored for outlier detection in the domain of fake news detection, addressing a critical gap in existing resources.
- We provide baseline experiments to facilitate rigorous evaluation and comparison of graph learning and LLM-based approaches.
- We release the dataset and code to the public, fostering further research in robust graph outlier detection and trustworthy AI.

2 Related Work

In this section, we review related datasets to our work. We compare TAGFN with existing datasets in Table 1 along the following aspects: presence of raw **Text** attributes, task of **Outlier** detection, scale 1M+ nodes/rows (**Large**), inclusion of **Graph** structure, and availability of **Time** information.

2.1 Text-Attributed Graph Datasets

Text-attributed graph (TAG) datasets have become a cornerstone for advancing research at the intersection of graph learning and LLMs. The recent surge in interest has led to the development of a diverse array of benchmarks. Yan et al. (2023) introduce CS-TAG, a diverse and large-scale suite of benchmark datasets for TAGs, and establish standardized evaluation protocols. Recognizing the importance of temporal dynamics, Zhang et al. (2024) introduced DTGB, a large-scale dynamic TAG dataset. Li et al. (2025c) further proposed TEG-DB, which incorporates both node and edge textual attributes. Despite these advances, there is no real-world TAG dataset for outlier detection.

2.2 Fake News Detection Datasets

Despite the proliferation of fake news detection datasets, most existing resources primarily focus on news content and basic metadata. Wang (2017) release is early fake new detection dataset LIAR, which includes PolitiFact-annotated short statements with rich meta-data (truthfulness, speaker, context, party affiliation, etc.) Nakamura et al. (2019) introduces Fakeddit, a multimodal fake news dataset of Reddit posts with paired text, images, and some metadata. FakeNewsNet (Shu et al., 2020) extends the LIAR by assembling multi-dimensional social media-based data, integrating full news content, social context, and spatiotemporal diffusion patterns. UPFD (Dou et al., 2021) further integrate user profiles and propagation graphs, enabling the study of social dynamics in fake news dissemination. However, UPFD lacks raw textual attributes, limiting its flexibility for LLM-based fake news detection. Despite recent advances in LLMs and outlier detection—such as AD-LLM Yang et al., 2024 and NLP-ADBench Li et al., 2024—existing approaches remain primarily focused on textual content. To bridge the gap, TAGFN offer not only news content, but also the propagation graph structure and user historical posts content, providing a more holistic view for fake news detection.

3 TAGFN

TAGFN is a text-attributed graph (TAG) dataset for graph level outlier detection in the domain of fake news detection. We present three subsets of varying scales: Politifact, Gossicop, and Fakeddit. Table 2 summarizes their statistics, including the number of nodes, edges, and graphs; the average graph size (in number of nodes); fake news ratio; and split size (train/validation/test) in graph count.

Table 2: Statistics of the three subsets of TAGFN.

	Politifact	**Gossipcop**	**Fakeddit**
# Nodes	41,054	314,262	7,249,803
# Edges	40,740	308,798	6,683,699
# Graphs	314	5,464	566,104
Avg. Size	131	58	13
Fake (%)	50.0	50.0	59.6
Train	62	1,092	467,538
Validation	31	546	49,186
Test	221	3,826	49,380

3.1 Problem Definition

While the dataset is in the domain of fake news detection, we formally define the general task of TAG outlier detection as follows:

Definition 1 (Text-Attributed Graph Outlier Detection)**.** *Let* $\mathbb{G} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ *denote a collection of* N *text-attributed graphs. Each graph* $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i, \mathbf{X}_i, \mathcal{T}_i)$ *consists of a set of nodes* $\mathcal{V}_i$*, a set of edges* $\mathcal{E}_i \subseteq \mathcal{V}_i \times \mathcal{V}_i$*, textual node attributes* $\mathbf{X}_i$*, and optional node associated timestamp* $\mathcal{T}_i$*. Each graph is annotated with a binary label* $y_i \in \{0, 1\}$*, indicating whether the graph is an outlier (*$y_i = 1$*) or not (*$y_i = 0$*). The objective of the task is to learn a function* $f : \mathcal{G} \to \{0, 1\}$ *that predicts the binary label* $\hat{y}_i$ *for each graph* $\mathcal{G}_i$*.*

3.2 Dataset Construction

To support outlier detection on TAG, we construct three fake news detection (sub-)datasets in the same format as `torch_geometric.data.Dataset`[1], based on existing datasets. For Politifact and Gossipcop, we follow (Shu et al., 2020) and (Dou et al., 2021), which jointly models news content and user interaction through propagation graphs. While the original datasets provided only preprocessed text embeddings via BERT or word2vec as node features, we retain the raw textual content of both the news articles and user historical posts, allowing for more flexibility in LLM-based methods. For Fakeddit, we adopt all samples in (Nakamura et al., 2019) as news content, and represent comment users[2] as child nodes in the graph. We filter out bot users[3] and remove the news that has no user comments. We mask out personal ID in the raw text on all subsets.

Figure 1 illustrates a toy example of a news propagation graph $\mathcal{G}$ in TAGFN. A detailed case study of a simple real instance is provided in Appendix A. The root node v_0 is the origin of the propagation graph (i.e., news itself), while the child nodes $(v_j, v_k) \in \mathcal{E}$ denote users involved in the propagation. Each node in the graph is attributed with text: the root node is attributed with the original news content $\mathbf{x}_0$, child nodes (users) are attributed with historical user posts $\mathbf{x}_j, j > 0$. To constrain the text length, we limit each user to their 200 most recent posts. Ground-truth outlier labels, indicating whether the news/graph is fake (1) or real (0), are adopted from prior work. Additionally, we include the Unix timestamp for each node to capture temporal information. Preliminary experiments indicate naively put the Unix timestamp into the LLM prompt does not significantly affect detection performance. We also provide public train, validation, and test splits consistent with (Shu et al., 2020) and (Nakamura et al., 2019).

4 Experiments

In this section, we benchmark the performance of current methods on TAGFN.

4.1 Prompting vs. Embedding

We start from evaluating different levels of supervision with LLMs.

- *Zero-shot inference*: standard **Zero-Shot** prompting and Chain-of-Thought **Reasoning** (Wei et al., 2022);
- *Few-shot in-context learning* (ICL; Brown et al., 2020), prompting the LLM with a few labeled examples per class, including **One-Shot**, **Two-Shot**, and **Three-Shot**;
- *Supervised learning*: training a graph neural network (GNN) on LLM-based embeddings, denoted as **Emb+GNN**, following UPFD (Dou et al., 2021).

Details of our prompt design for graph data are provided in Appendix B. For a fair comparison, we adopt Qwen3-8B (Yang et al., 2025) for prompting, and use Qwen3-Embedding-8B (Zhang et al., 2025) as the embedding model. We implement the GNN with GraphSAGE (Hamilton et al., 2017). We evaluate all the performance on test set of each

[1] https://pytorch-geometric.readthedocs.io/en/latest/generated/torch_geometric.data.Dataset.html

[2] https://pushshift.io

[3] https://botrank.pastimes.eu

Table 3: Performance across supervision levels (%).

Method	Politifact		Gossipcop		Fakeddit	
	ACC	F1	ACC	F1	ACC	F1
Zero-Shot	51.13	67.66	50.37	66.74	60.22	75.06
Reasoning	69.68	72.20	58.05	50.17	58.04	72.15
One-Shot	78.28	78.76	65.92	66.50	60.20	75.08
Two-Shot	69.23	74.44	58.29	42.01	60.22	75.10
Three-Shot	56.11	68.20	56.01	41.77	60.35	75.16
Emb+GNN	**84.16**	**83.72**	**96.71**	**96.75**	**84.93**	**87.99**

Table 4: Different LLMs on Politifact (%).

LLM	Zero-Shot		One-Shot	
	ACC	F1	ACC	F1
Qwen3-8B (2025)	51.13	67.66	78.28	78.76
Llama-3.1-8B (2024)	50.23	66.87	63.80	57.89
GPT-4.1-nano (2023)	51.58	67.87	57.47	69.68
GPT-4.1-mini (2023)	72.85	76.56	83.26	83.11
GPT-4.1 (2023)	**84.62**	**84.40**	**85.52**	**84.16**
O4-mini (2024)	79.64	80.35	81.90	81.98

subset, and sample the few-shot examples from validation set. The performance comparison of accuracy (ACC) and F1 score (F1) is shown in Table 3. From the table, we have three key findings:

1. In-context learning and reasoning help. We observe a substantial improvement in LLM accuracy on Politifact, rising from 51.13 to 78.28 with only one-shot examples. Moreover, even without any labeled examples, LLM reasoning also improves the accuracy to 69.68. A similar trend is observed on Gossicop, though not on Fakeddit. We hypothesize that this is due to the lower overlap between the pretraining data of Qwen3-8B and Fakeddit compared to other two subsets.

2. Supervised learning remains effective. Emb+GNN consistently outperforms Qwen3-8B-based methods across all three subsets. The performance gap is particularly pronounced on larger datasets, highlighting the importance of abundant supervision for effective fake news detection. Furthermore, when compared to the performance of BERT embeddings with GraphSAGE reported in Dou et al. (2021), the results using LLM-based embeddings are comparable, suggesting that the performance bottleneck lies not in the embeddings.

3. Two-shot and three-shot learning degrade with longer context. On Politifact and Gossicop, performance declines as the number of in-context examples increases. This degradation is not observed on Fakeddit, which has a smaller average graph size. We attribute this phenomenon to the increased context length, which may exceed the effective attention capacity of the LLM.

4.2 Performance of Different LLMs

Additionally, we benchmark the performance of various LLMs on Politifact in both zero-shot and one-shot settings, as summarized in Table 4. The results from GPT-4.1-nano to GPT-4.1 indicate that performance generally improves with increasing model size. Notably, zero-shot GPT-4.1 already surpasses the supervised Qwen3-Embedding-8B with GraphSAGE. In addition, providing just one example per class (one-shot ICL) yields a substantial boost for smaller models.

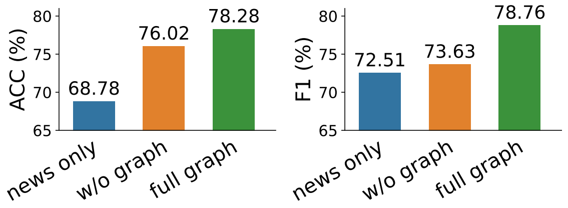

Figure 2: Ablation study of one-shot ICL on Politifact.

4.3 Ablation Study

We further conduct an ablation study on Politifact to show the importance of text-attributed graph. We consider two variants: **w/o graph**, which removes graph structure while retaining the new content and user posts in the prompt, and **news only**, which includes only the news content in the prompt. The results, shown in Figure 2, indicate that the full graph yields the best performance. Removing the graph structure or user posts leads to a noticeable drop in both accuracy and F1 score, highlighting the importance of each component in TAG.

5 Conclusion

To address the prevailing challenge of scarce real-world datasets, we introduce TAGFN, a text-attributed graph dataset designed for outlier detection, specifically fake news detection. TAGFN comprises of three subsets of varying scales, enabling comprehensive benchmarking. We further evaluate a range of methods on TAGFN to establish baseline performance. We hope this paper can provide valuable insights and facilitate future advancements in graph language models and trustworthy AI.

Limitations

As discussed in Section 3.2, our experiments on TAGFN are currently limited to static graphs, without considering the timestamp on each node. While preliminary experiments suggest that naive approaches to utilizing timestamps are ineffective, we reserve the exploration of more sophisticated temporal modeling techniques for future work. Furthermore, due to resource constraints and budgetary limitations, our evaluation does not include larger open-source models with over 10B parameters, as well as other prominent LLM families such as the Claude, Gemini, Mistral, and Deepseek.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and 1 others. 2024. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2):42–61.

Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, and Jiliang Tang. 2023. Label-free node classification on graphs with large language models (llms). *arXiv preprint arXiv:2310.04668*.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yanshu Li, Hongyang He, Yi Cao, Qisen Cheng, Xiang Fu, and Ruixiang Tang. 2025a. M2iv: Towards efficient and fine-grained multimodal in-context learning in large vision-language models. *arXiv preprint arXiv:2504.04633*.

Yanshu Li, Tian Yun, Jianjiang Yang, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025b. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. *arXiv preprint arXiv:2505.17098*.

Yuangang Li, Jiaqi Li, Zhuo Xiao, Tiankai Yang, Yi Nian, Xiyang Hu, and Yue Zhao. 2024. Nlp-adbench: Nlp anomaly detection benchmark. *arXiv preprint arXiv:2412.04784*.

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

Zhuofeng Li, Zixing Gou, Xiangnan Zhang, Zhongyuan Liu, Sirui Li, Yuntong Hu, Chen Ling, Zheng Zhang, and Liang Zhao. 2025c. Teg-db: a comprehensive dataset and benchmark of textual-edge graphs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Kay Liu, Yingtong Dou, Yue Zhao, Xueying Ding, Xiyang Hu, Ruitong Zhang, Kaize Ding, Canyu Chen, Hao Peng, Kai Shu, and 1 others. 2022. Bond: Benchmarking unsupervised outlier node detection on static attributed graphs. *Advances in Neural Information Processing Systems*, 35:27021–27035.

Kay Liu, Hengrui Zhang, Ziqing Hu, Fangxin Wang, and Philip S Yu. 2023. Data augmentation for supervised graph outlier detection with latent diffusion models. *arXiv preprint arXiv:2312.17679*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

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.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2258–2268.

Haoyan Xu, Zhengtao Yao, Ziyi Wang, Zhan Cheng, Xiyang Hu, Mengyuan Li, and Yue Zhao. 2025a. Graph synthetic out-of-distribution exposure with large language models. *arXiv preprint arXiv:2504.21198*.

Haoyan Xu, Zhengtao Yao, Xuzhi Zhang, Ziyi Wang, Langzhou He, Yushun Dong, Philip S Yu, Mengyuan Li, and Yue Zhao. 2025b. Glip-ood: Zero-shot graph ood detection with graph foundation model. *arXiv preprint arXiv:2504.21186*.

Hao Yan, Chaozhuo Li, Ruosong Long, Chao Yan, Jianan Zhao, Wenwen Zhuang, Jun Yin, Peiyan Zhang, Weihao Han, Hao Sun, and 1 others. 2023. A comprehensive study on text-attributed graphs: Benchmarking and rethinking. *Advances in Neural Information Processing Systems*, 36:17238–17264.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. 2021. Graphformers: Gnn-nested transformers for representation learning on textual graph. *Advances in Neural Information Processing Systems*, 34:28798–28810.

Tiankai Yang, Yi Nian, Shawn Li, Ruiyao Xu, Yuangang Li, Jiaqi Li, Zhuo Xiao, Xiyang Hu, Ryan Rossi, Kaize Ding, and 1 others. 2024. Ad-llm: Benchmarking large language models for anomaly detection. *arXiv preprint arXiv:2412.11142*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Jiasheng Zhang, Jialin Chen, Menglin Yang, Aosong Feng, Shuang Liang, Jie Shao, and Rex Ying. 2024. Dtgb: A comprehensive benchmark for dynamic text-attributed graphs. In *Advances in Neural Information Processing Systems*, volume 37, pages 91405–91429. Curran Associates, Inc.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yun Zhu, Haizhou Shi, Xiaotang Wang, Yongchao Liu, Yaoke Wang, Boci Peng, Chuntao Hong, and Siliang Tang. 2025. Graphclip: Enhancing transferability in graph foundation models for text-attributed graphs. In *Proceedings of the ACM on Web Conference 2025*, pages 2183–2197.

A Data Instance

To provide a case study on TAGFN, we illustrate an instance of news propagation graph from Politifact in Figure 3. This graph has 4 nodes (from Node 0 to Node 3) and 3 edges. The corresponding raw textual attributes for each node are as follows:

```
Node 0 (News Content): Based on the
    Monthly Treasury Statement for
    August and the Daily Treasury
    Statements for September. CBO
    estimates that the federal budget
    deficit was about $1.30 trillion in
    fiscal year 2011, approximately the
    same dollar amount as the shortfall
    recorded in 2010. The 2011 deficit
    was equal to 8.6 percent of gross
    domestic product, CBO estimates,
    down from 8.9 percent in 2010 and
    10.0 percent in 2009, but greater
    than in any other year since 1945.
    The estimated 2011 total reflects
    the shift of some payments from
    fiscal year 2012 into fiscal year
    2011 (that is, from October to
    September, because October 1 fell on
     a weekend); without that shift, the
     deficit in 2011 would have been $1
    .27 trillion. CBO's deficit estimate
     is based on data from the Daily
    Treasury Statements; the Treasury
    Department will report the actual
    deficit for fiscal year 2011 later
    this month.

Node 1 (User Post): Teri and I wish you
    a Merry Christmas!  Wishing peace
    and joy to my friends and neighbors
    in the Jewish community on this
    first night of Hanukkah ...

Node 2 (User Post): RT @user: LATE
    BREAKING: This morning the FBI
    arrested a member of the Cincinnati
    city council for accepting bribe
    money in exch ...

Node 3 (User Post): .@user There they go
     again, with superficial over #
    Substance. One wd suggest they Try
    to compare their er ...
```

B Prompt Design

To enable LLM-based fake news detection on text-attributed graphs, we design a structured prompt to encode both the news content and the associated propagation graph for LLMs, following graph

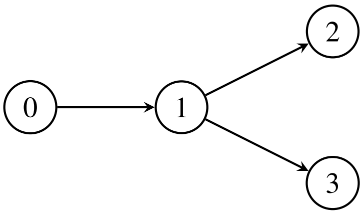

Figure 3: The graph structure of the data instance.

prompt in (Fatemi et al., 2024). The prompt includes a system prompt and a user prompt.

B.1 System Prompt

The system prompt sets the context for the LLM, describing the task and the format of the input. The following system prompt is used for experiments:

System Prompt

```
You are a fake news detection assistant analyzing news propagation graph on social networks.

You will be provided with:
- the content of a news (corresponding to the root node in the propagation graph),
- user posts (each corresponding to a subsequent node in the propagation graph),
- the structure of the propagation graph. The edges indicate the propagation relationships.

Based on the content and graph structure, your task is to determine whether the news is ‘Real‘ or ‘Fake‘.

Output: respond with only the fake news classification label: ‘Real‘ or ‘Fake‘.
```

B.2 User Prompt

The user prompt encodes the instance to be classified, including the news content, user posts, and the graph structure. For few-shot in-context learning, a few labeled examples are included in the same format, each followed by the correct output label (‘Real‘ or ‘Fake‘). In experiments, to fit the prompt into the context window, we restrict the post content of each user to 500 characters and the maximum number of users to 30. Below is a demostration of the prompt provided to the LLM:

User Prompt

```
EXAMPLES:

Input:
Node 0 (NEWS): <news content>
Node 1 (USER POST): <user post>
Node 2 (USER POST): <user post>

Graph Structure:
Node 0 propagate to Node 1,
Node 0 propagate to Node 2

Output: Real

Input:
Node 0 (NEWS): <news content>
Node 1 (USER POST): <user post>
Node 2 (USER POST): <user post>

Graph Structure:
Node 0 propagate to Node 1,
Node 1 propagate to Node 2

Output: Fake
END OF EXAMPLES. Classify the following news:

Input:
Node 0 (NEWS): <news content>
Node 1 (USER POST): <user post>
Node 2 (USER POST): <user post>

Graph Structure:
Node 0 propagate to Node 1,
Node 1 propagate to Node 2
```