Detecting LLM-Generated Spam Reviews by Integrating Graph Neural Network and Language Model Embeddings

Anonymous ACL submission

Abstract

Detecting spam reviews has drawn much attention for years. Many efforts have been dedicated to detecting deceptive spam reviews, accumulating rich literature and plentiful effective practices. However, the recent rapid development of large language models (LLMs) brings new challenges to this area. Fraudsters could misuse LLMs to write highly authentic and misleading fake reviews. To detect such harmful contents, we formulate the detection as a node classification task on the constructed review graph and employ the graph neural network (GNN) to handle the users' behavior. More specifically, we seamlessly integrate gated graph transformers with the language model to embed the review texts where previously engineered features summarized by fraud experts are insufficient. The Experiments show that this integration in our method FraudSquad turns out to be effective on two created LLM-attacked and two human-attacked spam review datasets, outperforming state-of-the-art detection methods. Moreover, FraudSquad achieves a modest model size and requires very few training labels, making defending the spam review attack more practical.

1 Introduction

Online reviews significantly influence consumer decisions and business reputations (Duma et al., 2023), but the prevalence of spam reviews, which are deceptive reviews meant to mislead consumers, poses a serious challenge (Jindal and Liu, 2008; Andresini et al., 2022). They cause financial losses to both honest merchants and consumers. Detecting spam reviews is essential to protect online consumers and maintain the credibility of review platforms like Amazon and Taobao.

However, this detection task remains a challenging area. Despite numerous studies (Liu et al., 2021a; Xiang et al., 2023; Duan et al., 2024), a comprehensive solution is yet to be achieved, highlighting the need for continuous research and innovation.

Moreover, recent developments in generative large language models have made it easier to generate sophisticated misleading fake reviews, making detection even more important (Zellers et al., 2019). In the era of large language models, fraudsters may exploit the public information including product metadata and genuine review texts to generate fake content that is difficult to distinguish from that written by genuine users.

From the detector side, many current fraud detection methods often overlook the rich linguistic features embedded in review texts, which can provide crucial insights for identifying sophisticated fake reviews. They rely on engineered features and focus on the relationships between reviews, reviewers, products and other entities. Many methods model the problem using graph-based techniques (Tian et al., 2015; Hooi et al., 2016) and recently graph neural networks (GNNs) have been effective in capturing complex interactions within review networks (Zhang et al., 2020a).

Therefore, we propose a hybrid approach, FraudSquad, that integrates language model embeddings and graph neural networks to enhance the detection of spam reviews. FraudSquad enhances the node embeddings by text embedding from a pre-trained language model on the constructed review graph and then applying the gated graph transformers for fraud classification. In this way, it leverages both linguistic and relational data. By carefully designing and choosing the sub-modules, FraudSquad achieves a modest and concise model, providing accurate results without complex feature engineering as in many existing works.

To evaluate the detection efficiency against LLMs-generated spam review, we create two LLM-spammed datasets as there is a lack of data. The generation pipeline on the real Amazon dataset considers various information that a fraudster could

input to an LLM-based chatting assistant: product meta information, reference review text and output requirements. We employ GPT-4o to evaluate the generated texts, finding that these fake reviews are highly persuasive for potential customers. Moreover, the generation satisfies the output requirements with high speed for automatic attack, increasing the urgency of accurate detectors.

Luckily, our method FraudSquad turns out to accurately detect these LLM-generated spam reviews, outperforming state-of-the-art fraud detectors by at least 10% in terms of various metrics and achieving overall metric scores of more than 90% with only 1% annotated labels. In addition, FraudSquad is also significantly more effective on two human-written spam review datasets. Experiments verify that advanced text embedding and graph structure are indispensable for accurate detection, saving the labor of maintaining complexly engineered features.

2 Methodology

2.1 Problem Formulation

To utilize both the review text and user behavior, we model the spam review detection problem as a node classification task on the review graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, T)$. Here $\mathcal{V} = \{v_1, \dots, v_N\}$ is the node set representing N reviews. $\mathcal{E}$ denotes the edge set containing the relationships between review nodes. The text of the review node $v_i \in \mathcal{V}$ is denoted as T_i which contains multiple tokens $(T_{i1}, T_{i2}, \dots)$. Some nodes are associated with a class $y_i \in \{0, 1\}$ where 0 represents normal and 1 represents fraud (spam). There is typically a very small set of labeled nodes for training, while most nodes remain unlabeled. Our task is to predict the fraud nodes from unlabeled ones on the graph.

2.2 FraudSquad Detector

Our detection model FraudSquad contains three main modules, review graph construction, LM-enhanced node embeddings, gated graph transformers and MLP layers for classification.

Review graph construction. The review graph is constructed to represent the behavior of users by incorporating various edge types. In a typical review scenario such as Amazon, we connect review nodes through three types of relations as prescribed by previous works (Dou et al., 2020; Liu et al., 2021b): (1) connecting review pairs written by the same *user*; (2) connecting reviews pairs towards the same *product* with the same *star* rating; (3) connecting reviews pairs towards the same *product* in the same *month*. For other scenarios such as question-answering portals where the answer gives a review of the product mentioned in the question, we could also construct a graph as follows as an example: (1) connecting QA pairs under the same *question*; (2) connecting connects QA pairs whose *questions* are given the same user within the same *month*; (3) connecting connects QA pairs whose *answers* are given the same user within the same *month*.

LM enhanced node embeddings. The initial node embeddings on the constructed review graph are obtained as follows. We first input each review text T_i to a pre-trained language model (LM) with frozen weights and get its embedding in a latent space, denoted as X_i.

Next, we derive a trainable risk embedding Z_i from the labels so that label propagation can be executed at the same time with a graph neural network. As pointed out by Shi et al. (2021), unifying label propagation with feature (text embedding here) propagation in the graph neural network is beneficial. Therefore, we treat the label as a special categorial feature. The size of the risk embedding vocabulary is hence three (normal, fraud and unknown) and the embedding size is set to equal the text embedding size. To avoid label leakage, in each training batch, the labels of the training nodes are all masked as unknown and these nodes could only aggregate the risk embeddings from the neighboring nodes.

Finally, the initial node embeddings combine both the text embedding and the risk embedding with trainable weights β_1, β_2:

$$H_i = X_i + \text{PReLU}(X_i\beta_1 + Z_i\beta_2)\beta_3. \quad (1)$$

Gated graph transformers. Now we input the initial node embeddings H_i into gated graph transformer layers with multi-headed attention. The graph transformer architecture (Dwivedi and Bresson, 2020) has three fundamental embedding vectors for every node v_i, the Query (Q) Value (V) and Key (K):

$$\begin{aligned} Q_{is} &= H_i W^s_{query}, \\ V_{is} &= H_i W^s_{value}, \\ K_{is} &= H_i W^s_{key}. \end{aligned} \quad (2)$$

Here $W^s_{query}, W^s_{value}, W^s_{key}$ are the weights for the s-th attention head and we have $s = 1, \dots, S$

heads. Next, consider a specific node v_i and denote its neighboring node set as $\mathcal{N}_i$. We compute the attention coefficients between the centering node v_i and every neighboring node $v_j \in \mathcal{N}_i$:

$$\alpha^s_{i,j} = \frac{\exp(Q^T_{is} K_{js})}{\sum_{v_j \in \mathcal{N}_i} \exp(Q^T_{is} K_{js})}, \tag{3}$$

which measures the similarity between the Query embedding of node v_i and the Key embedding of node v_j. Then the attention coefficients determine the magnitude of the contribution of neighboring nodes' Value embedding to the update of the centering node:

$$\tilde{H}^s_i = \sum_{v_j \in \mathcal{N}_i} \alpha^s_{ij} V_{js}. \tag{4}$$

After concatenating all heads, we have:

$$\tilde{H}_i = \text{Concat}(\tilde{H}^1_i, \cdots, \tilde{H}^S_i). \tag{5}$$

Finally, we add a shortcut from the linearly transformed input of the layer O_i using a gate operation the same as Xiang et al. (2023):

$$\begin{aligned} O_i &= H_i \beta_3, \\ \text{gate}_i &= \text{Sigmoid}(\text{Concat}(O_i, \tilde{H}_i, O_i - \tilde{H}_i)\beta_4), \\ \hat{H}_i &= \text{gate}_i O_i + (1 - \text{gate}_i)\tilde{H}_i. \end{aligned} \tag{6}$$

$\hat{H}_i$ is the final output of one gated graph transform layer.

MLP classification. After $L = 2$ such gated graph transformation layers, an MLP Layer outputs the fraud probability. The entire model FraudSquad is trained using labeled nodes with binary cross-entropy loss and optimized with the Adam optimizer.

3 Synthesizing Training and Evaluation Datasets

Since there is a lack of public datasets to evaluate the detection against LLM-generated review spam, we synthesize such datasets here.

3.1 Generating Review Spam Texts using LLMs

Suppose a fraudster wants to post review spam to a target product. The goal of the attack is to generate reviews as specific as possible to the target product and also indistinguishable from real ones. So, we consider the following information that a fraudster could input to an LLM-based chatting assistant. (See Table 7 in the appendix for detailed prompts.) All the product meta information and reference review texts could be extracted from e-commerce websites such as Amazon.

- Product meta information: its name, category and official description given by the selling store.
- Reference review texts: the genuine reviews that the product receives whose sentiment may be positive or negative.
- Output requirements: positive or negative, max word number, asking for diversified contents with different lengths and styles, and output format that each review should be in a new paragraph.

We apply this pipeline to a `Amazon` dataset. This dataset is derived from a large-scale Amazon Review Dataset (Hou et al., 2024). Specifically, we select reviews from the year 2022, covering eight categories: baby products, video games, software, musical instruments, appliances, all beauty, health and personal care, and digital music. The `Amazon` dataset comprises 7,617 products and 86,758 reviews in total.

We focus on generating positive (five-star) review texts for products with the lowest average rating stars and the fewest number of reviews. Given that more than 75% of products have an average rating higher than 4.3 on a five-star scale, it is reasonable to assume that products below this threshold aim to enhance their reputation. We select 500 products and generate five positive review texts for each, with a maximum output length of 100 words. The reference review text is derived from the first review of the target product.

Two LLMs are used for the generation: (1) Llama3-8B[1] and (2) Qwen2-72B-Instruct (Yang et al., 2024). Both models are open-source and implemented using the Ollama[2] framework. They operate within the same container environment, utilizing 8 CPU cores, 5000 MB of memory, and 3 NVIDIA GeForce RTX 4090 GPUs, each with 24 GB of memory.

3.2 Evaluating Generated Texts.

First, LLMs generate fake review texts that are highly persuasive for potential customers and look

[1] https://github.com/meta-llama/llama3
[2] https://github.com/ollama/ollama

Table 1: Statistics of fake review text generation outputs using LLMs.

LLM	Outputted/Required	Max. #Words	Ave. #Words	Speed
Llama3	2488/2500	94	56.5 ± 7.4	0.5 s/review
Qwen2	2500/2500	133	54.5 ± 8.7	2.8 s/review

much like being posted by normal users. Here we employ GPT-4o deployed on Azure to evaluate if the generated review texts by LLMs appear positive, detailed, convincing, human and influential from a potential customer's perspective and give a five-grade score (See Table 8 in the appendix for detailed prompts.). For comparison, we also randomly choose 2500 five-star human-written review texts from the Amazon dataset and evaluate them.

Figure 1 shows the evaluation results by GPT-4o. It turns out that both Llama3 and Qwen2 achieve nearly full scores with a low variation in all evaluating factors. Furthermore, the LLMs-generated review texts even outperform human-written review texts, especially in being convincing and influential. All these show that LLMs-generated review texts could greatly influence potential customers by giving misleading information.

Second, these generated review texts satisfy the output requirements in the meantime. Table 1 shows the statistics of generation results. The outputted review number is extremely close to the required number, indicating that both two LLMs follow the output format strictly well so the review texts could be extracted automatically at a high rate (more than 99.5%). Besides, the max word number requirement is also satisfied in most cases where Llama3 is absolutely under the requirement (100).

To check the diversity requirement, we randomly select some review texts [3], finding that they differ a lot in expression and focus on various aspects of the product.

3.3 Injecting Generated Review Spam to the Amazon dataset

Now we assume that fraudsters compromise some existing users who have posted normal reviews in the original Amazon dataset, using these accounts to inject the review texts generated by LLMs. This creates highly challenging, camouflaged detection scenarios (Dou et al., 2020). The compromised users are sampled with a probability proportional to the number of reviews they have written. Each

[3]All generated data will be published once accepted.

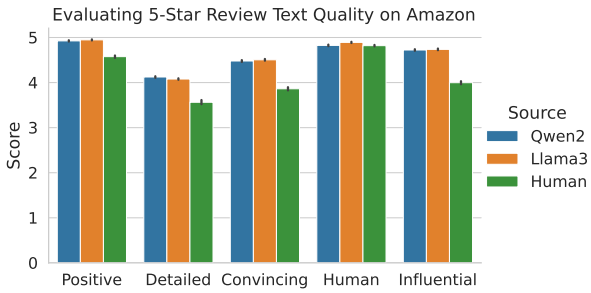


Figure 1: Evaluating five-star review texts generated by LLMs and human on Amazon by GPT-4o.

compromised user is tasked with writing two fake reviews for the target product, both of which are given a five-star rating. The time of these reviews is randomly sampled to occur at any hour within five days of the first review of the target product.

4 Experiment Setup

Datasets. We establish four datasets for evaluation. Besides the LLM-spammed dataset synthesized in Sec.3, we also employ two datasets containing human-written review spam. Both of them have ground-truth labels. Yelp (Rayana and Akoglu, 2015) is a public dataset containing Yelp reviews for hotels in Chicago and filtered(spam)/recommended(normal) labels provided by Yelp. Liu et al. (2017) collect the CQA dataset on a Chinese community question-answering website in 2015. It has normal question-answering pairs and manipulated content posted by massive organized crowd-sourcing workers to mislead common users, such as brand promotion campaigns which is much similar to the review spam. So we regard the manipulated contents as the goal for detection. The detailed data statistics are in Table 2.

Compared methods. TThe proposed hybrid model FraudSquad was compared with several baselines to demonstrate its effectiveness in fraud review detection. Baselines 1-3 are the general classification methods and baselines 4-7 are more recent GNN-based fraud detection methods. These models as well as FraudSquad are implemented

Table 2: Experiment dataset statistics

Dataset	Nodes	Edges	Fraud
Amazon-Llama3	89,192	4,139,448	2.8%
Amazon-Qwen2	89,192	4,140,166	2.8%
Yelp	5,854	141,123	13.3%
CQA	133,317	66,272,741	34.2%

with PyTorch. We also use DGL (Wang et al., 2020) to implement the graph neural networks in FraudSquad.

(1) MLP: multi-layer perceptron with two hidden layers that receives numerical features as inputs.

(2) RNN: recurrent neural network that receives texts as inputs.

(3) GAT (Veličković et al., 2018): graph attention network that utilize a graph with node features for node classification.

(4) CARE-GNN (Dou et al., 2020): graph neural network with enhanced aggregation modules against fraud camouflages.

(5)PC-GNN (Liu et al., 2021b): graph neural network that addresses the label-imbalance problem in fraud detection.

(6) GTAN (Xiang et al., 2023): gated temporal attention network for credit-card and other fraud detection tasks.

(7) DGA-GNN (Duan et al., 2024): dynamic grouping aggregation graph neural network for fraud detection.

These models as well as FraudSquad are implemented with PyTorch We also use DGL (Wang et al., 2020) to implement the graph neural networks in FraudSquad.

For models that require numerical node features including MLP and GNN-based comparing baselines, we feed them with the previous engineered fraud features (Dou et al., 2020) (See Table 9 in the appendix). We employ these features for Yelp and two LLM-spammed Amazon datasets. The CQA dataset contains different features indicative of fraud (Liu et al., 2017), such as if the answer is posted by a master user. And we also apply the engineered features with a slight change to the CQA dataset.

Training setting. The FraudSquad training parameters are as follows. The model employs the BERT-base-uncased (Devlin et al., 2019) as the text embedding large language model. The gated graph neural network uses a hidden dimension of 100 and the number of attention heads is set to 3. The maximum number of training epochs is 50. When dividing the datasets into training, validation and testing, we consider a barely supervised setting as acquiring labels in fraud detection is typically difficult (Yu et al., 2024). The portions of the three parts are set as 1%-9%-90%. Since the CQA dataset contains more nodes, the training set is scaled down to 0.1%-9.9%-90%. All methods including FraudSquad run on one GPU with 48GB of memory.

Evaluation metrics. Evaluation metrics used include precision, recall, AP (average precision), and AUC (the area under the ROC Curve). After the detection method outputs the probabilities being fraud, we predict the nodes with highest probabilities as fraud and then compute the precision and recall scores. The routine of a real fraud detection system usually involves manual check. Therefore, it is important to give an appropriate number of suspicious node candidates with high precision and recall. The top ratios are set as 5%, 30%, 3% and 3% respectively for the datasets in Table 2.

5 Experiment Results

5.1 Detecting LLMs-generated review spam.

Table 3 shows the performance against LLMs-generated review spam. The highest metric values are in **bold** and the second best ones are underlined.

All the metric scores of our method FraudSquad are higher than 0.90. It indicates that the LLM-generated review spam attack could still be accurately detected, greatly relieving the potential risks of misusing LLMs to conduct review spam attack.

Moreover, the two attacking LLMs show differences in evading detection, especially graph-based detectors. Qwen2 is better at evading graph based detectors than Llama3. The metric scores including precision, recall and AP are hardly above 0.5 when detecting Qwen2 generated review spam. On the contrary, the first three metric scores could reach nearly 0.8 when detecting Llama3.

Besides, FraudSquad outperforms baselines to a large extent especially in terms of precision, recall and AP.Though other methods including RNN and DGA-GNN could have a relatively high AUC, their precisions and recalls are significantly low. It means that the top suspicious review nodes they predict are not accurate. This also indicates that in this very imbalanced setting, a thorough evalua-

Table 3: Detection performance against LLMs-generated review spam on the Amazon dataset.

Attacking LLM	Qwen2				Llama3			
Method	Precision	Recall	AP	AUC	Precision	Recall	AP	AUC
MLP	0.3526	0.3780	0.3126	0.7937	0.0000	0.0000	0.0287	0.5564
RNN	0.4814	0.5601	0.5726	0.9641	0.6886	0.5746	0.6775	0.9407
GAT	0.4348	0.4662	0.3869	0.8609	0.2147	0.2308	0.1350	0.7745
CARE-GNN	0.4448	0.4768	0.4283	0.8969	0.4568	0.4911	0.4628	0.9563
PC-GNN	0.3198	0.3428	0.2313	0.8343	0.4286	0.4607	0.4651	0.9469
GTAN	0.4514	0.4840	0.4101	0.9000	0.6287	0.6759	0.7040	0.9691
DGA-GNN	0.3641	0.4265	0.3769	0.8905	0.7823	0.8085	0.7934	0.9330
FraudSquad	**0.9236**	**0.9902**	**0.9957**	**0.9998**	**0.9099**	**0.9781**	**0.9905**	**0.9994**

Table 4: Detection performance against human-written review spam.

Dataset	Yelp				CQA			
Method	Precision	Recall	AP	AUC	Precision	Recall	AP	AUC
MLP	0.2776	0.1043	0.2147	0.6530	0.6358	0.5579	0.6005	0.7742
RNN	0.1768	0.1200	0.1559	0.5580	0.3274	0.0531	0.3867	0.5853
GAT	0.2395	0.0900	0.1830	0.6003	0.5741	0.5037	0.6250	0.7200
CARE-GNN	0.2877	0.1080	0.1971	0.6031	0.6847	0.6008	0.6963	0.7878
PC-GNN	0.2814	0.1057	0.1909	0.5902	0.6696	0.5876	0.6950	0.7868
GTAN	0.2205	0.0829	0.1718	0.5795	0.6963	0.6110	0.7016	0.7981
DGA-GNN	0.2069	0.0943	0.1683	0.5601	0.4670	0.5684	0.4717	0.6680
FraudSquad	**0.5057**	**0.1900**	**0.3602**	**0.6870**	**0.9991**	**0.8767**	**0.9902**	**0.9943**

tion is necessary to investigate the model's performance.

5.2 Detecting human-written review spam

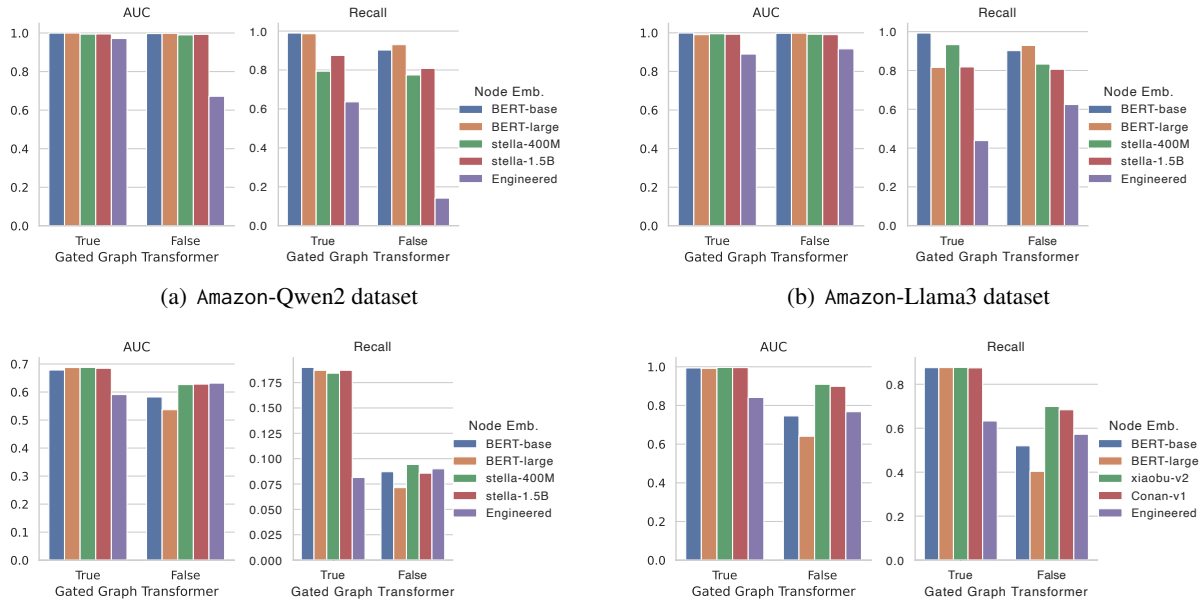
Table 4 shows the overall detection performance against human-written review spam. FraudSquad still achieves the highest metric scores on both datasets.

In addition, we notice that detection on Yelp dataset is harder than on CQA dataset. The metric scores of detection on Yelp are no higher than 0.7 and often below 0.2. On the contrary, metric scores of detection on CQA is mostly more than 0.6 and FraudSquad could achieve more than 0.90. One reason is that the review spam on CQA dataset originally has multiple expressive features (such as the grades of the asker and answerer) and the spamming behaviors are more coherent (Liu et al., 2017). So feeding these features into MLP is enough to have a reasonable performance, that is, precision and recall above 0.5. On the other hand, the ground-truth label on Yelp dataset is annotated by more complex mechanisms.

5.3 Benefit of LM enhanced node embeddings

Here we compare the detection performance of using language model enhanced node embeddings with engineered-feature based node embeddings. We consider four LMs here. Two BERT (Devlin et al., 2019) models and two more recent text embedding LLMs which rank top in the classification category of Massive Text Embedding Benchmark (Muennighoff et al., 2022). Since CQA is a Chinese dataset, two Chinese embedding models are chosen here.

First of all, all four LM enhanced node embeddings yield much better detection performance than engineered-feature based node embeddings across four datasets as shown in Figure 2, especially on the two LLMs generated review spam datasets. On the real Yelp and CQA datasets, the engineered-feature based embeddings may perform competitively when no graph information is used (gated graph transformer is False), but still not the best.



(a) Amazon-Qwen2 dataset

(b) Amazon-Llama3 dataset

(c) Yelp dataset

(d) CQA dataset

Figure 2: Benefit of LLM enhanced node embeddings and gated graph transformer.

Comparing the default BERT models with more recent embedding LLMs, it is obvious that BERT has a strong classification capability with a modest model size, appearing to be a good choice. This may be the recent LLM embedding models support a broader spectrum of tasks. Besides, we also notice that when there is no graph information on the CQA dataset, the two Chinese LLMs perform better than BERT, indicating the importance of training corpora.

5.4 Benefit of gated graph transformer.

Though text embedding produced by LLMs could be directly input into a linear classifier and show good accuracies, utilizing the node relationships by gated graph transformer still benefits a lot. In Figure 2, we conduct the ablation study by replacing the gated graph transformer (True) with one MLP layer (False). In the latter case, no graph information is utilized. The recall metric scores demonstrate significant difference on four datasets, especially for detecting human-written review spam.

5.5 Limited utility of engineered features

To further investigate the superiority of LLM text embeddings, we consider concatenating the initial node embeddings with engineered fraud features derived from raw data by previous domain experts (See Table 9 in the appendix) before the next gated graph transformer layers. Tables 5-6 show the gaps between the using engineered features and not. It turns out that engineered features have an insignificant influence on the detection metric scores, especially in detecting LLMs-generated review spam. The gap is sometimes positive and sometimes negative. Therefore, we suggest the practitioners could solely rely on the LLM text embeddings if they find the engineered features time-consuming to develop and maintain.

6 Related Work

The detection of fraud reviews has been extensively studied covering various aspects, from proposing meaningful features to efficient detection methods. And the misinformation detection in other areas also gives insights that could be borrowed.

Engineered fraud features. Numerous studies have been carried out to analyze the characteristics of fraudulent behavior and especially review spam (Li et al., 2011; Lim et al., 2010). Table 9 in the appendix lists the engineered features for review spam detection (Rayana and Akoglu, 2015), which serve as the basis of fraud detection. Integrating linguistic features with behavioral data, for example, has been shown to improve performance over using either input alone. Approaches that bridge review networks and metadata, utilizing both textual and relational data, demonstrate the potential for more accurate and robust detection (Rayana and Akoglu, 2015; Wang et al., 2017). These features are still

Table 5: Studying the engineered features on the Amazon dataset against LLMs-generated review spam.

Attacking LLM	Qwen2				Llama3			
Engineered features	Precision	Recall	AP	AUC	Precision	Recall	AP	AUC
True	**0.9282**	**0.9951**	**0.9977**	**0.9999**	0.8567	0.9210	0.9163	0.9951
False	0.9269	0.9938	0.9971	0.9997	**0.9182**	**0.9871**	**0.9934**	**0.9998**
Gap	0.0013	0.0013	0.0006	0.0002	-0.0615	-0.0661	-0.0771	-0.0047

Table 6: Studying the engineered features against human-written review spam.

Dataset	Yelp				CQA			
Engineered features	Precision	Recall	AP	AUC	Precision	Recall	AP	AUC
True	**0.5323**	**0.2000**	**0.3665**	0.6894	0.9983	0.8760	0.9892	0.9937
False	0.5133	0.1929	0.3587	**0.7020**	**0.9994**	**0.8770**	**0.9896**	**0.9939**
Gap	0.0190	0.0071	0.0078	-0.0126	-0.0011	-0.0010	-0.0004	-0.0002

used widely in recent spam detection works (Dou et al., 2020; Duan et al., 2024).

Linguistic-based detection. These methods focus on analyzing the textual content of reviews to identify deceptive patterns. Common techniques include sentiment analysis, syntactic analysis, and lexical feature extraction. For instance, sentiment and psycholinguistic features have been incorporated to achieve higher detection accuracy, though these models often struggle with sophisticated fraudsters who mimic genuine review characteristics (Ott et al., 2011; Li et al., 2014).

Graph-based detection. Graph-based approaches leverage the relationships between reviews, reviewers, and products. Techniques such as Graph Convolutional Networks (GCNs) have shown promise in capturing complex interactions (Liu et al., 2018). Recent advancements in graph neural networks have addressed more challenges like class imbalance (Liu et al., 2021b), categorical feature (Xiang et al., 2023; Duan et al., 2024) and camouflage (Zeng and Tang, 2021; Yu et al., 2024) in fraud detection tasks. For instance, spectral analysis has been incorporated to enhance detection capabilities (Zhang et al., 2020b; Xu et al., 2023). Furthermore, methods like CARE-GNN and its improved variant, RLC-GNN, address issues of relation and feature camouflage, demonstrating significant improvements in fraud detection performance (Zeng and Tang, 2021). Additionally, (Xiang et al., 2023; Duan et al., 2024) effectively handles the non-additive categorical features.

Misinformation detection in other areas. Large language models pose great threats in a variety of misinformation generation, including fake news, rumors, hallucination and persuasion (Wu et al., 2024; Xu et al., 2024). For example, the adversaries could generate targeted propaganda that closely mimics the style of real news (Zellers et al., 2019). DECOR is a novel approach to the fake news detection problem, which leverages the large language model to embed the news texts and then train a GNN-based detector with a social graph refinement component (Wu and Hooi, 2023). Its effectiveness and efficiency are inspiring to our review spam detection.

7 Conclusion

In this work, we study spam review detection in the presence of large language models. We enhance existing graph neural network detectors by integrating language models to embed review texts with gated graph transformers. Due to a lack of data, we synthesize two LLM-spammed datasets by simulating a scenario where a fraudster interacts with an LLM-based chatting assistant and finds that the generated reviews appear highly authentic, highlighting the need for accurate detectors. The hybrid approach FraudSquad we propose proves effective in detecting both LLM-generated and human-written review spam, demonstrating the importance of integrating these two techniques to improve detection accuracy and counter the evolving nature of fraudulent activities.

Limitations

During data synthesizing, we mainly focus on generating positive review texts as it is reasonable to boost a low-reputation product for a merchant. However, in some other scenarios, some people may directly post negative review texts to hurt the reputation of a popular product of their competitors or simply ask for money to delete those malicious reviews. This is due to the concern such tools may have more unwanted results when published. This may be a future direction with more thorough thoughts

Besides, the datasets in the detection experiment involve two human-written spam review datasets, one from Yelp and one from a Chinese question-answering portal. They are not as recent as the Amazon dataset. This is because such review-related fraud detection datasets are usually maintained within a company for privacy reasons. We hope some more datasets may be open to research access to test the detection performance.

Ethics Statement

The adversarial nature of spam review detection involves threat modeling. Though the generation pipeline in the evaluation may increase the risk of dual-use, our latter detection model FraudSquad shows extraordinary performance with high precisions and recalls, requiring very few annotated labels. We believe this result would prevent such LLM misuse in advance and have a positive impact for the common good.

Furthermore, our detection model FraudSquad is not 100% accurate. Some honest reviews may be falsely identified as deceptive spam, for example, an individual user who uses the large language model to help write an honest review. Therefore, we expect human efforts after the detection. Some widely adopted routines, including manually checking the detected suspicious contents and responding to user feedback, should help recover justice and fairness.

When writing this paper, we used AI assistants ChatGPT and DeepSeek purely to polish the language of the paper. The data used in this work are properly used and cited solely for academic purposes.

References

Giuseppina Andresini, Andrea Iovine, Roberto Gasbarro, Marco Lomolino, Marco Degemmis, and Annalisa Appice. 2022. Review spam detection using multi-view deep learning combining content and behavioral features. In *itaDATA*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Yingtong Dou, Zhiwei Liu, Li Sun, Yutong Deng, Hao Peng, and Philip S. Yu. 2020. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*.

Mingjiang Duan, Tongya Zheng, Yang Gao, Gang Wang, Zunlei Feng, and Xinyu Wang. 2024. Dga-gnn: Dynamic grouping aggregation gnn for fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10):11820–11828.

Ramadhani Ally Duma, Zhendong Niu, Ally S. Nyamawe, Jude Tchaye-Kondi, and Abdulganiyu Abdu Yusuf. 2023. A deep hybrid model for fake review detection by jointly leveraging review text, overall ratings, and aspect ratings. *Soft Computing*, 27:6281–6296.

Vijay Prakash Dwivedi and Xavier Bresson. 2020. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*.

Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. 2016. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 895–904. ACM.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM.

Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2011. Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 2488–2493. AAAI Press.

Fei Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2014. Towards a holistic approach to detect spam reviews in online review platforms. *Proceedings of the 23rd*

International Conference on World Wide Web, pages 459–470.

Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, page 939–948, New York, NY, USA. Association for Computing Machinery.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. 2021a. Pick and choose: A gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, pages 3168–3177.

Yang Liu, Xiang Ao, Zidi Qin, Jianfeng Chi, Jinghua Feng, Hao Yang, and Qing He. 2021b. Pick and choose: A gnn-based imbalanced learning approach for fraud detection. In *Proceedings of the Web Conference 2021*, pages 3168–3177.

Yuli Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2017. Detecting collusive spamming activities in community question answering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1073–1082, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li, and Le Song. 2018. Heterogeneous graph neural networks for malicious account detection. In *CIKM*, pages 2077–2085. ACM.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 985–994.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. 2021. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*.

Tian Tian, Jun Zhu, Fen Xia, Xin Zhuang, and Tong Zhang. 2015. Crowd fraud detection in internet advertising. In *WWW*, pages 1100–1110.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the Sixth International Conference on Learning Representations*.

Jian Wang, Shuhua Feng, Bing Liu, and Yuming Li. 2017. Using a hybrid content-based and behavior-based featuring approach in fake review detection. In *Proceedings of the 2017 International Conference on Information Systems*, pages 849–861.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2020. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *Preprint*, arXiv:1909.01315.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 3367–3378, New York, NY, USA. Association for Computing Machinery.

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

Sheng Xiang, Mingzhi Zhu, Dawei Cheng, Enxia Li, Ruihui Zhao, Yi Ouyang, Ling Chen, and Yefeng Zheng. 2023. Semi-supervised credit card fraud detection via attribute-driven graph representation. In *The Annual AAAI Conference on Artificial Intelligence*.

Fan Xu, Nan Wang, Hao Wu, Xuezhi Wen, Xibin Zhao, and Hai Wan. 2023. Revisiting graph-based fraud detection in sight of heterophily and spectrum. *arXiv preprint arXiv:2312.06441*.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin,

Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hang Yu, Zhengyang Liu, and Xiangfeng Luo. 2024. Barely supervised learning for graph-based fraud detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16548–16557.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32.

Yufan Zeng and Jiashan Tang. 2021. Rlc-gnn: An improved deep architecture for spatial-based graph neural network with application to fraud detection. *Applied Sciences*, 11(12):5656.

Shijie Zhang, Hongzhi Yin, Tong Chen, Quoc Viet Nguyen Hung, Zi Huang, and Lizhen Cui. 2020a. Gcn-based user representation learning for unifying robust recommendation and fraudster detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 689–698, New York, NY, USA. Association for Computing Machinery.

Yu Zhang, Pang-Ning Tan, and Ying Ding. 2020b. Fraud review detection using graph convolutional networks. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pages 2773–2781. ACM.

A Detailed Prompts in the Case Study

B Engineered fraud features

Table 7: Prompt of generating the review texts using LLMs

Role	Message
User	Hello, I need your help to write reviews to a product [product title] on Amazon in the category of [product categoty].
User	The official description of the product given by the store is as follows. [product description]
User	Besides, I will give you a review to this product for reference. [reference review text]
User	Now please output [review number] [positive/negative] reviews. Each review contains no more than [max word] words. Please write diversified reviews as if they were written by different customers, for example, with different length and styles. Start with another paragraph for each review and begin with Review 1. etc. Thanks a lot!!

Table 8: Prompt of evaluating the five-star review texts by GPT-4o

Role	Message
System	You are a helpful assistant and know a lot about e-commerce on Amazon, especially about how the reviews influence potential customers.
User	Please first read a review about the product titled [product title] in the category of [product categoty].
User	[evaluating review].
User	Now could you help me evaluating the influence of the given review to a potential customer on Amazon in the following four aspects.
User	Will the user feel the review is positive?
User	Will the user feel the review contains useful details?
User	Will the user feel the review is convincing?
User	Will the user feel the review is written by a normal user?
User	Will the user be more willing to buy the product after reading the review?
User	For each question, please first answer an integer ranging from 1 (totally no) to 5 (totally yes) and then give a brief reason within 20 words.

Table 9: Feature engineering for review spam detection(Rayana and Akoglu, 2015; Dou et al., 2020)

Review	Review length in words
	Ratio of ALL-capitals words
	Ratio of capital letters
	Ratio of first-person pronouns ('I', 'me'...)
	Ratio of subjective words (by sentiWordNet)
	Ratio of objective words (by sentiWordNet)
	Ratio of exclamation sentences containing '!'
	Rating deviation within the product
	Thresholded rating deviation
	If rating is extreme (1 or 5 star)
	If review is user's sole review
	If review is the product's early review
User & Product	Average review length in words
	Average text similarity in 1 and 2-grams.
	Max text similarity in in 1 and 2-grams.
	Max. number of reviews in a day
	Max. number of reviews in a month
	Ratio of positive reviews (4-5 star)
	Ratio of negative reviews (1-2 star)
	Average rating
	Entropy of rating distribution
	Entropy of temporal gaps

Table 10: Generated fake review text examples by Llama3 targeting at products in the Amazon dataset.

Category	Review text
Musical instruments	I'm not a professional musician, just a hobbyist who loves to play for fun. I was blown away by the quality of these strings - they brought out the best in my guitar and made it sound like a completely different instrument. The price is very reasonable too.
All beauty	I'm thrilled with the results I've gotten from Scruples Quickseal Detangling Conditioner! As someone who color-treats their hair regularly, it's amazing to find a product that not only detangles but also helps maintain my color vibrancy. This stuff has become a regular in my beauty routine - can't recommend it enough!
Baby products	I was blown away by this pet gate! As a busy pet parent, I need solutions that work quickly and easily. This magic gate exceeded my expectations - it's a breeze to set up, sturdy, and does exactly what I need it to do: keep my furry friend safe indoors. Plus, the design is sleek and modern, making it a great addition to my home.
Video games	What a fantastic product! I was skeptical at first, but the 60-day pre-paid time card has given me the freedom to play without interruption. The in-game content is top-notch, and the community is so engaging. A must-have for any Star Wars fan like myself
Software	What a treat! M&M'S Adventure has become my go-to distraction when I need a break. The puzzles are cleverly designed, and the visuals are top-notch. I love how the game keeps me on my toes with new levels and characters to discover. It's also great that it's free, with no hidden costs or in-app purchases to worry about. A must-have for any puzzle game fan!
Appliances	I was skeptical at first, but this light has become an essential part of my DIY projects! The adjustable clip-on feature makes it easy to position the light exactly where I need it. And the different lighting settings are great for working on various materials - from wood to metal. It's also very bright and the battery life is excellent.

Table 11: Generated fake review text examples by Qwen2 targeting at products in the Amazon dataset.

Category	Review text
All beauty	Absolutely in love! I've never felt my skin so soft after using an exfoliating product. The 100% PURE Cocoa Coffee Body Scrub is a game-changer. It feels like a spa treatment at home. The scent is divine, a warm and comforting aroma of cocoa that's not overpowering. The combination of sea salts and nourishing oils leaves my skin smooth as silk without any irritation. Highly recommend this for a luxurious pampering session!
Video Games	Absolutely thrilled with my purchase! The $40 Xbox Gift Card was delivered instantly to my inbox, allowing me to download the latest games without any hassle. It's perfect for gamers who prefer digital copies and it's also great when you want to treat yourself or a friend to some new adventures.
Baby products	As a new mom, finding the right products for my baby has been essential. The KAKIBLIN Pillow is truly a lifesaver! Its soft fabric combined with organic pp cotton ensures that my little angel gets the best comfort possible. Even during long drives, he seems relaxed and happy. The fact that it's machine washable adds to its convenience. A must-have for parents!
Software	Absolutely thrilled with my new Rii Wired Gaming Controller! It's super responsive and the dual-vibration/turbo feature really enhances my gaming experience on PC and Android. Love that it's plug and play, so easy to set up.
Health and personal care	I've been using this shaver for a couple of weeks now, and I'm impressed by how well it performs every time. The cordless design adds so much flexibility in my routine. It's also great that the battery holds up really well between charges. A must-have for on-the-go grooming!
Appliances	I was quite skeptical about purchasing parts online, but the W10197428 combination kit proved me wrong. It's an exact fit for my Whirlpool fridge. Installation was straightforward, and my refrigerator hasn't skipped a beat since then. The customer service provided by this store is also top-notch, responsive, and helpful. I highly recommend it!