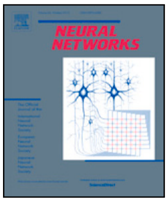
Neural Networks 187 (2025) 107377

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet



Full Length Article

Enhancing text-centric fake news detection via external knowledge distillation from LLMs

Xueqin Chen [a], Xiaoyu Huang [b], Qiang Gao [b,c,*], Li Huang [b,c], Guisong Liu [b,c]

[a] *Kash Institute of Electronics and Information Industry, Kashgar, China*
[b] *School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, China*
[c] *Engineering Research Center of Intelligent Finance, Ministry of Education, Chengdu, China*

ARTICLE INFO

Keywords:
Fake news detection
SLM
External knowledge
LLM
Feature alignment
Knowledge distillation

ABSTRACT

Fake news poses a significant threat to society, making the automatic and accurate detection of fake news an urgent task. Various detection cues have been explored in extensive research, with news text content shown to be indispensable as it directly reflects the creator's intent. Existing paradigms for developing text-centric methods, i.e., small language model (SLM)-based, external knowledge-enhanced, and large language model (LLM)-based approaches, have achieved remarkable improvements. However, each of these paradigms still faces the following challenges: (1) the low generalization ability of SLM-based methods, due to their training on limited and specific knowledge; (2) the extensive retrieval operations required by external knowledge-enhanced methods, both during training and at the inference stage, leading to increased computational costs; and (3) LLMs are prone to hallucinations and less suited for factual reasoning. To address these challenges, we propose LEKD, which combines the strengths of SLMs, external knowledge, and LLMs to enhance text-centric fake news detection. Specifically, LEKD leverages the LLM to generate external knowledge as supplementary information for the training set only and introduces a graph-based semantic-aware feature alignment module to resolve knowledge contradictions, as well as an information bottleneck-based knowledge distillation module to ensure the implicit generation of these features during inference. Extensive experiments conducted on two datasets demonstrate the advantages of LEKD over the baselines.

1. Introduction

In the contemporary digital era, the emergence of online social platforms has accelerated the spread of information among people. This convenience of rapid information dissemination creates a breeding ground for fake news, posing real-world threats in many crucial aspects such as politics (Allcott & Gentzkow, 2017; van der Linden, Panagopoulos, & Roozenbeek, 2020), the economy (Arcuri, Gandolfi, & Russo, 2023), and public health (Rocha et al., 2021). Thus, accurately detecting fake news has become an urgent task for our society. Initially, fake news detection relied on crowdsourcing through fact-checking websites, such as PolitiFact,[1] FactCheck,[2] and Snopes.[3] With the advancement of machine learning technologies, automatic detectors have gradually entered the scene, particularly with the rise of deep learning-based methods.

Specifically, state-of-the-art deep learning-based methods primarily focus on capturing multimodal features from various inputs, such as news text (Ma et al., 2016; Yu, Liu, Wu, Wang, & Tan, 2017), images (Wang et al., 2018; Wu, Liu, Zhao, Wang & Zhang, 2024), and social context (Chen, Zhou, Trajcevski, Bonsangue, Marcello, 2022; Shu, Cui, Wang, Lee, & Liu, 2019). These features are then fused to create a comprehensive news representation, which is used for the final prediction. Moreover, among these multimodal data, news text content has proven to be an indispensable clue, as it reflects the writer's intentions and is directly linked to the credibility of the news (Zhou & Zafarani, 2020). In contrast, visual features may introduce ambiguity caused by image manipulation or repurposing (Chen et al., 2022; Dong et al., 2024), while social context features are often difficult to access due to privacy concerns and also carry inherent uncertainties (Zhang et al., 2024). Thus, in this work, we aim to enhance *text-centric fake news detection.*

* Corresponding author at: School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, China.
E-mail addresses: nedchen0728@gmail.com (X. Chen), 223081200023@smail.swufe.edu.cn (X. Huang), qianggao@swufe.edu.cn (Q. Gao), lihuang@swufe.edu.cn (L. Huang), gliu@swufe.edu.cn (G. Liu).

[1] http://www.politifact.com/
[2] https://www.factcheck.org/
[3] https://www.snopes.com/

https://doi.org/10.1016/j.neunet.2025.107377
Received 1 October 2024; Received in revised form 28 February 2025; Accepted 7 March 2025
Available online 17 March 2025

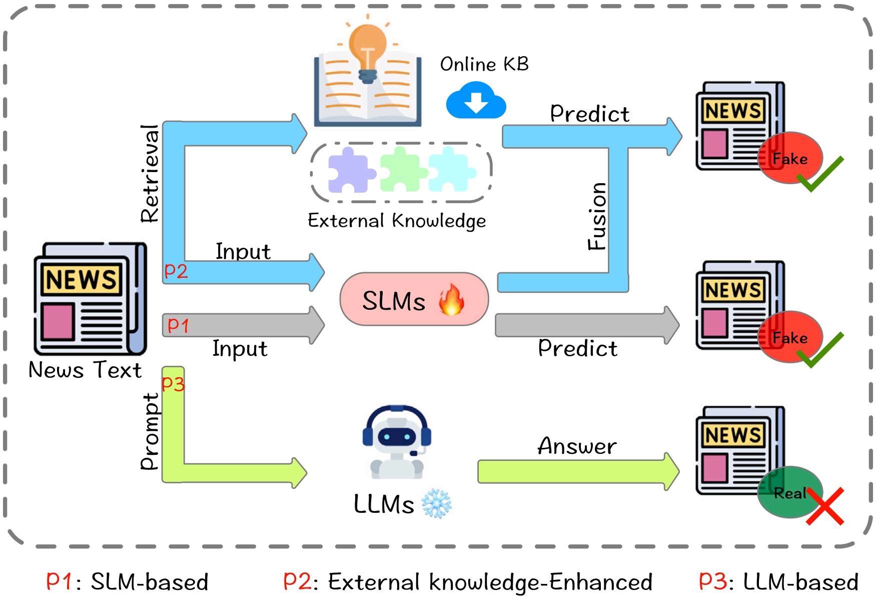


Fig. 1. Taxonomy of advanced text-centric fake news detection methods.

Existing common paradigms for text-centric methods can be broadly categorized into three, as illustrated in Fig. 1: (*P1*) A core paradigm is the small language model (SLM)-based approach, which leverages pretrained SLMs to extract textual features from news content, serving as cues for fake news detection. This backbone can be seamlessly integrated into diverse fake news detection models. Typical SLMs include BERT (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (Liu et al., 2019), etc. (*P2*) Following paradigm P1, several studies have explored retrieving evidential texts from online knowledge bases, such as knowledge graphs and Wikipedia, to supplement the factual content of original news. This external knowledge is then integrated with the original textual features of the news for final prediction. (*P3*) The advent of large language models (LLMs) introduces a novel paradigm that directly engages LLMs in reasoning about the truthfulness of given news through manually designed precise prompts.

Each of the three paradigms has its own advantages: P1 relies on pre-trained SLMs, not only offering a comprehensive understanding of news text, but also being easily fine-tuned for diverse downstream tasks and plugged into other models; P2 incorporates external knowledge, enriching data resources and enhancing the model's ability to understand the news, thereby improving detection performance; and P3 leverages the strong reasoning abilities and extensive world knowledge of LLMs, offering interpretability of its reasoning steps and accessibility for non-experts. However, they still face some challenges: (**C1**) Diverse SLMs are trained on different text corpora, leading to low generalization ability when handling out-of-knowledge news items (Sheng, Zhang, Cao, & Zhong, 2021). (**C2**) The acquisition of external knowledge requires extensive retrieval operations from knowledge bases during both the training and inference phases, leading to additional computing resources and time. Moreover, (**C3**) the choice of knowledge bases also causes the detection model to face the same generalization issues as SLM-based methods (Hu et al., 2024). Furthermore, (**C4**) LLMs face the hallucination problem, which can generate unintentional content, thereby influencing reasoning accuracy (Chen & Shu, 2024).

To address the above concerns, in this work, we propose LLM-based External Knowledge Distillation (LEKD) model, aimed at improving the basic SLM-based text-centric fake news detection model by introducing external knowledge and leveraging the natural language understanding and generation capabilities of LLMs. In a nutshell, our LEKD consists of five components: an LLM-based knowledge enrichment module, a pre-trained SLM-based knowledge encoder, a graph-based semantic-aware feature alignment module, a knowledge cross-distillation module, and an MLP-based detector. Specifically, at different stages of the model training lifecycle, i.e., preprocessing, training, and inference, different components of LEKD will be invoked. (S1) In the **preprocessing** phase, LEKD introduces an LLM-based knowledge enrichment module to generate external knowledge (background information and inconsistency points) for news items in the training set using simple, human-designed prompts. (S2) During model **training**, the original news text and generated external knowledge are fed into a pretrained SLM to produce text embeddings. Then, a graph-based semantic-aware feature alignment module is deployed: on one hand, to mitigate the hallucination problem in LLM-generated content, and on the other, to resolve contradictions between the original news and external knowledge. Subsequently, an information bottleneck-inspired knowledge cross-distillation module is introduced to distill external knowledge into two injectors, ensuring the implicit generation of external knowledge features during inference without requiring retrieval from LLMs. (S3) In the **inference** phase, only the original news text is used as input for LEKD. The fine-tuned SLM and two injectors are then called upon to generate the news text embedding and supplementary external knowledge features. Finally, these features are utilized as cues for prediction.

Overall, the main contributions of this work can be summarized as follows:

- We propose a novel text-centric fake news detection model, LEKD, built upon small language model-based methods and enhanced by seamlessly integrating the strengths of external knowledge-enhanced and large language model-based approaches. LEKD not only improves detection performance but also enhances model generalization, reduces reliance on retrieval operations, and avoids unreliable reasoning.
- The proposed model leverages an LLM to generate external knowledge, i.e., background information and inconsistency points. This external knowledge acquisition is performed only on the training set during the preprocessing phase. We ground in the information bottleneck theory and introduce a knowledge cross-distillation module to transfer latent external knowledge into knowledge distillation injectors, ensuring the implicit generation of external knowledge features during inference.
- We design a graph-based semantic-aware feature alignment module to align the representations of the original news text and complementary knowledge into the same semantic space. This

helps to alleviate hallucinations in LLM-generated external knowledge and resolve contradictions between the original news and the generated knowledge.
- We conducted extensive experiments on two real-world datasets, Weibo and GossipCop, using various evaluation metrics to demonstrate the detection effectiveness of the proposed LEKD compared to state-of-the-art baselines. Moreover, ablation studies and rich visualization results further emphasize the importance of each component and showcase LEKD's generalization ability.

The remainder of this paper is organized as follows. First, we provide a review of text-centric fake news detection methods in Section 2. Subsequently, we formalize the problem definition, followed by a detailed description of our proposed LEKD in Section 3. The experimental results demonstrating the advantages of LEKD are presented in Section 4. Finally, we conclude this study and outline potential future directions.

2. Related work

2.1. Fake news detection

Fake news detection aims to determine the truthfulness of the given news article. Existing research continues to enhance detection accuracy by incorporating various multimodal features, while news text content remains indispensable, serving as the primary clue for detecting fake news. In this section, we first briefly review recent text-centric methods, followed by a discussion of advanced efforts in LLM-based detectors.

Text-centric methods focus on manually extracting rich textural features or automatically learning representative text embeddings using deep learning technologies from a given news article. Then, these extracted features or learned embeddings are employed as key indicators for assessing the truthfulness of the news. Specifically, conventional methods manually define a set of handcrafted features based on characters, words, sentences, and documents within the news article, including lexical features (Castillo, Mendoza, & Poblete, 2011; Kwon, Cha, Jung, Chen, & Wang, 2013; Zhao, Resnick, & Mei, 2015), syntactic features (Castillo et al., 2011; Hassan, Qazvinian, & Radev, 2010; Ma, Lin, & Cao, 2017), topic features (Wu, Yang, & Zhu, 2015), etc. These selected features are then used as predictors and fed into discriminative machine learning algorithms, e.g., decision trees (Loh, 2011), naive Bayes (Bayes, 1968), and support vector machines (Hearst, Dumais, Osuna, Platt, & Scholkopf, 1998), for classification. However, on the one hand, the performance of these handcrafted feature-based methods heavily depends on the quality of the defined features. On the other hand, there is a lack of a standardized and systematic way to design general features across different platforms, and the manual process is also time-consuming.

Over the past decade, deep learning has revolutionized the field of natural language processing (NLP), driving many researchers to shift their focus toward utilizing various deep learning technicals to learn text embeddings for fake news detection (Chang, Li, & Duan, 2024; Chen et al., 2024) . Among the earliest approaches are a series of end-to-end methods based on recurrent neural networks (RNNs) or convolutional neural networks (CNN) (Raj & Meel, 2022). For example, Ma et al. (2016) proposed the first RNN-based to capture both temporal and textual features from posts related to the statement. Yu et al. (2017) introduced a CNN-based approach to enhance detection accuracy by extracting key features and modeling high-level interactions among these significant features. However, these RNN/CNN-based methods are inefficient at modeling long texts and only perform well in short-sentence prediction tasks. Subsequently, the emergence of the Transformer (Vaswani et al., 2017) ushered text-centric methods into a new generation, i.e., attention-based methods. Initially, researchers focused on incorporating attention mechanisms into their models to better capture word/sentence interactions, thereby enhancing the model's ability to learn text representations (Chen, Sui, Hu, & Gong, 2019; Shu et al., 2019; Trueman, J., P., & J., 2021). Later, the widespread use of pre-trained small language models (SLMs), such as BERT (Kaliyar, Goswami, & Narang, 2021; Sheng et al., 2022; Zhu et al., 2022), as feature extractors further boosted model performance. In addition, some works have explored enhancing textual feature extraction by incorporating novel techniques, such as adversarial learning (Ma, Gao, & Wong, 2019) and graph neural networks (Vaibhav, Mandyam, & Hovy, 2019).

Recently, researchers have noted that the limited available knowledge from original news text hinders the performance of existing methods. This has facilitated the development of external knowledge-enhanced methods (Ling, Chen, Lai, & Liu, 2024). These methods execute retrieval operations on knowledge bases (KBs), such as Freebase, Wikidata, DBpedia, and Google's Knowledge Graph, to acquire evidential information. Specifically, they retrieve relevant entities from the knowledge graph through a process that includes entity linking and entity content extraction. The news content, along with the extracted entities and their contexts, are then jointly used for prediction. For instance, work (Hu et al., 2021) compares contextual entity representations with KB-based entity representations using a comparison network to capture the consistency between the news content and the KB. Finally, the news representation, combined with the entity comparison features, is fed into a fake news classifier. Dun et al. (Dun, Tu, Chen, Hou, & Yuan, 2021) propose a Knowledge-aware Attention Network (KAN), which aligns news entities with knowledge graph entities and employs attention mechanisms to capture the importance of both semantic-level and knowledge-level representations, thereby enhancing the detection of fake news. The introduction of external knowledge not only provides complementary information to improve model performance but also offers a degree of interpretability (Popat, Mukherjee, Yates, & Weikum, 2018; Schlicht, Sezerer, Tekir, Han, & Boukhers, 2021; Sheng et al., 2021; Wu, Rao, Yang, Wang, & Nazir, 2020). However, different news domains rely on distinct KBs, and the extensive retrieval operations occur not only during training but also during the inference phase, leading to increased resource demands and higher time costs. In addition, these methods involve entity abstraction at the sentence level, rather than considering the news article as a whole at the document level. Our work builds upon the concept of external knowledge-enhanced methods, but we retrieve external knowledge based on the entire news text content, and this acquisition step is only performed on the training set during the preprocessing stage. Our approach significantly reduces the need for retrieval operations while still allowing for the implicit injection of external knowledge during the inference phase.

Large language models (LLMs), once introduced, aroused interest from both industry and academia across various application domains due to their remarkable abilities in understanding and generating language. Unsurprisingly, several LLM-based works have also been proposed in the field of fake news detection. Given that LLMs can mimic human writing styles to generate text based on input prompts, some works have focused on distinguishing between human-generated fake news and LLM-generated fake news (Chen & Shu, 2023; Huang, McKeown, Nakov, Choi, & Ji, 2023; Lucas et al., 2023; Pan et al., 2023), demonstrating that LLMs can be effective misinformation generators. Other works directly utilize LLMs as fact-checkers (Guan, Dodge, Wadden, Huang, & Peng, 2024; Pan et al., 2023; Pelrine et al., 2023; Zhang & Gao, 2023), leveraging their vast word knowledge and strong reasoning abilities. Chen et al. (Chen & Shu, 2024) conducted a systematic analysis of LLMs in both combating and generating misinformation, exhibiting that, in contrast to fact-checking, LLMs show stronger capabilities in generating misinformation. Similarly, the latest work (Hu et al., 2024) highlighted the hallucinations of LLMs and their lack of a reliable factual basis, making them unsuitable for direct use as fake news detectors. However, their strong commonsense reasoning capabilities can be leveraged to elicit external knowledge, improving previous SLM-based methods. Our LEKD is similar to the work of Hu, Sheng, et al. (2024), but focuses on leveraging the power of LLMs in text summarization and commonsense reasoning, using zero-shot prompts to generate relevant external knowledge as complementary information for the training set, rather than exploring various prompt strategies for generating rational information.

2.2. Knowledge distillation

Knowledge distillation (KD) is first introduced by Hinton et al. (Hinton, Vinyals, & Dean, 2015), which is used for model compression by transferring knowledge from a well-trained, larger teacher model to a smaller student model. This approach enables the student model to achieve performance comparable to the teacher model while being significantly more efficient in terms of computation and memory usage. The classic KD paradigm typically adopts an offline style, where the teacher model is pre-trained separately and then supervises the student model's training, often leading to increased complexity and resource demands. To address these challenges, online distillation (Zhang, Xiang, Hospedales, & Lu, 2018) was introduced, enabling simultaneous training of teacher and student models, thus streamlining the process. Additionally, self-distillation (Zhang et al., 2019) emerged as an approach where a single model improves itself using its own outputs, eliminating the need for a separate teacher model. Furthermore, advanced KD methods have been developed to address zero-shot (Micaelli & Storkey, 2019), multi-modal (Gupta, Hoffman, & Malik, 2016), and multi-task (Zhang & Peng, 2018) scenarios.

Recently, in the context of KD, researchers have begun exploring distlling task-related information rather than focusing solely on architecture compression (Aghli & Ribeiro, 2021; Wang et al., 2019). Specifically, these approaches leverage information bottleneck (IB) theory (Tishby, Pereira, & Bialek, 2001) to distill and retain the most relevant information for the given task (Dai, Zhu, Guo, & Wipf, 2018; Srivastava, Dutta, Gupta, Agarwal, & AP, 2021). Dai et al. (2018) proposed the first IB-based KD method, known as Variational Information Bottleneck (VIB), which leverages the IB principle through a variational bound to reduce redundancy across layers. Subsequently, Ayush et al (Srivastava et al., 2021) extended VIB to sequential models, achieving high compression rates and faster inference speeds for action recognition tasks. More recently, Tian et al. (Tian, Zhang, Lin, Qu, Xie, & Ma, 2021) advanced the IB framework by introducing Variational Self-Distillation (VSD), which extends Variational Cross-Distillation (VCD) to multi-view learning. These works provide a strong theoretical foundation for our approach. Building on this, our LEKD leverages VIB theory to distill external knowledge into a lightweight, offline injector optimized for efficient inference.

3. Methodology

In this section, we introduce our proposed rumor detector – LEKD, which mainly consists of five modules: (1) **LLM-based Knowledge Enrichment** enriches the original news text by gathering external knowledge from multiple perspectives through an LLM; (2) **Knowledge Encoder** utilizes pre-trained SLM to extract diverse and complex content features and generate representations for news articles and external knowledge; (3) **Graph-based Semantic-aware Feature Alignment** aims to alleviate hallucinations and ambiguity in external knowledge compared to the original news text; (4) **Knowledge Cross-Distillation** is built upon information bottleneck theory, transferring external knowledge into two knowledge distillation injectors to ensure the generation of latent external knowledge during inference; and (5) **Detection** outputs the news credibility based on the news text representation and latent external knowledge. The overall framework is shown in Fig. 2. In the following subsections, we will describe each component in detail.

3.1. Problem formulation

Suppose we have a news article dataset $\mathcal{D}_{\text{news}}$ on hand, each news article $D_i \in \mathcal{D}_{\text{news}}$ is represented by a tuple (t_i, y_i), where t_i denotes the text content of D_i, and $y_i \in \{0, 1\}$ is its corresponding credibility label. Here, 1 and 0 indicate that the news is fake and true, respectively. As a result, the dataset can be expressed as $\mathcal{D}_{\text{news}} = \{(t_1, y_1), (t_2, y_2), \ldots, (t_N, y_N)\}$, where N is the total number of news articles in the dataset. Ultimately, the fundamental issue tackled in this study can be defined as:

Text-centric Fake news detection: Given a news article D_i, the text-centric fake news detection task aims to learn a function f_θ that extracts comprehensive textual features from t_i and uses them as cues to predict the corresponding credibility label for D_i. This can be further formulated as:

$$\hat{y}_i = \text{MLP}(f_\theta(t_i)). \tag{1}$$

where $\text{MLP}(\cdot)$ represents the multi-layer perceptron used for label prediction, and $\hat{y}_i$ is the predicted label for D_i.

3.2. LLM-based knowledge enrichment

As we reviewed in Section 2, external knowledge from reliable KBs can enrich the original news text content, thereby improving the performance of standard text-centric methods. However, these methods still face the following challenges: (1) ***Unrestricted KB selection***. There is no specific rule for selecting the most appropriate KB from various options, since the choice heavily depends on human expertise and specific domain requirements. For instance, misinformation in the healthcare domain requires knowledge from a medical knowledge graph (Cui et al., 2020), while political news needs insights from political-related KBs (Hu et al., 2021). (2) ***Superfluous time cost***. During both training and inference, each news article requires the retrieval of entity descriptions and interactions from the KBs (Dun et al., 2021; Hu et al., 2021). These extensive and redundant retrieval operations incur additional time costs due to response delays.

In this work, to address the two limitations mentioned above, we leverage LLMs to generate external knowledge as supplementary data for the training set, prior to training the text-centric fake news detection model. LLMs are well known for their extensive *world knowledge* and *strong reasoning capabilities*, which have revolutionized the artificial general intelligence (AGI) field. They are widely applied to a variety of complex NLP tasks, such as summarization (Onan & Alhumyani, 2024) and question answering (Arefeen, Debnath, & Chakradhar, 2024). By virtue of the parametric knowledge embedded in LLMs, we utilize zero-shot prompting technical (Si et al., 2022) to ask an LLM (specifically GPT-3.5 Turbo[4] in our implementation) to generate reasonable and informative supplementary knowledge for a given news article.

Zero-shot Prompting. For a given news D_i, we first construct the prompt p_i for D_i in a uniform format, where p_i only consists of the predefined task description template $\mathcal{T}$ and the given news text content t_i, and is formulated as:

$$p_i = (\mathcal{T}, t_i). \tag{2}$$

In our implementation, the template $\mathcal{T}$ is tailored to match the writing language of the news article D_i. And the prompt template is designed to specifically inquire about two types of knowledge: **Commonsense knowledge** and **Background knowledge**, both of which are commonly explored in existing works and have been demonstrated to be effective in fake news detection (Cui et al., 2020; Hu et al., 2021; Sun, Zhang, Zheng, & Ma, 2022; Wu, Yu, Chen & Zhou, 2024). Instead of directly asking LLM to summarize the commonsense knowledge within the news, we prompt it to reason about points that deviate from commonsense, adopting an inverse approach. In detail, Table 1 outlines

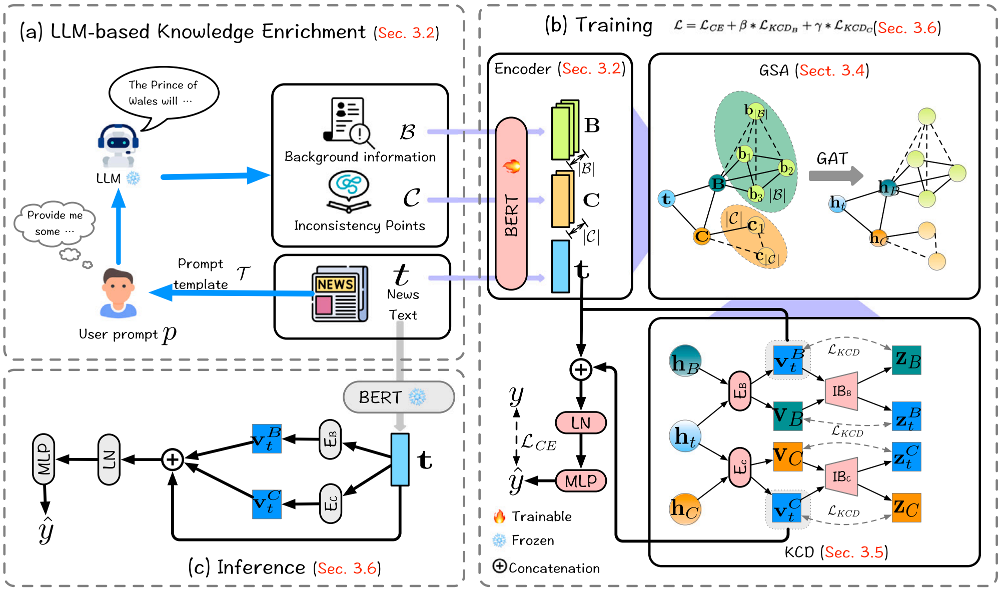


Fig. 2. Overall framework of LEKD.

Table 1
Prompt templates for LLM.

Prompt template in English	Prompt template in Chinese

Given the following news article [t_i], please provide relevant background information and identify points where the news does not conform to common sense. Present your answers concisely, using the following JSON format:
{
"BIs": [
"b1",
"b2",
...
],
"IPs": [
"c1",
"c2",
...
]
}
where BIs and IPs stand for background information and knowledge contradictory to common sense for the given news, respectively. Ensure that your response is in valid JSON format.
	请针对以下新闻文章 [t_i] 提供相关背景信息,
	并指出该新闻不符合常识的点。请简洁地以以下 JSON 格式呈现您的答案:
	{
	"BIs": [
	"b1",
	"b2",
	...
	],
	"IPs": [
	"c1",
	"c2",
	...
	]
	}
	其中 **BIs** 和 **IPs**
	分别代表了新闻的背景信息以及不符合常识的内容。请确保您的回复为标准的
	JSON 格式。

the prompt template used in this work for both English and Chinese news.

Next, we input the prompt into the LLM, which then returns a series of background information (BIs) and inconsistency points (IPs). This process is formulated as follows:

$$(\mathcal{B}_i, \mathcal{C}_i) = \text{LLM}(p_i), \tag{3}$$

where $\mathcal{B}_i = \{b_1, b_2, \dots, b_{|\mathcal{B}_i|}\}$ denotes the background knowledge, and $\mathcal{C}_i = \{c_1, c_2, \dots, c_{|\mathcal{C}_i|}\}$ represents the points of inconsistency with commonsense. $|\mathcal{B}_i|$ and $|\mathcal{C}_i|$ are the numbers of BIs and IPs obtained for news t_i, respectively. Note that, in practice, LLMs may provide varying numbers of BIs ($|\mathcal{B}_i|$) and IP ($|\mathcal{C}_i|$) for different news articles. To this end, for each news article D_i in $\mathcal{D}$, we incorporate its $\mathcal{B}_i$ and $\mathcal{C}_i$ with t_i and y_i to form the enhanced data quadruple $D_i^{\text{aug}} = (t_i, \mathcal{B}_i, \mathcal{C}_i, y_i)$, and the news dataset after knowledge enrichment is denoted as $\mathcal{D}_{\text{aug}} = \{D_1^{\text{aug}}, \dots, D_i^{\text{aug}}, \dots, D_N^{\text{aug}}\}$.

Discussion. Compared with the conventional external knowledge acquisition methods, ours is (1) *One-Time Interaction with KB*. Our external knowledge acquisition process is implemented during the preprocessing phase, requiring only a one-time interaction with the LLM. In contrast, conventional methods involve extensive retrieval operations from KBs during both model training and inference. (2) *Unnecessary to specify KB*. Our method relies on the superior parametric knowledge of LLMs. Regardless of which LLM is selected, they are all trained on diverse KBs and encompass information from various domains.

3.3. Knowledge encoder

After enriching the training set, we utilize a pre-trained small language model to encode the text content into latent representations. Specifically, we chose BERT (Devlin et al., 2019) as the text encoder

[4] https://platform.openai.com/docs/models/gpt-3-5-turbo

backbone in this work. For a given augmented news article D_i^{aug}, we input each element of the triple $(t_i, \mathcal{B}_i, \mathcal{C}_i)$ into BERT separately, i.e,

$$\mathbf{t}_i = \mathrm{BERT}(t_i), \tag{4}$$

$$\mathbf{B}_i = \mathrm{BERT}(\mathcal{B}_i), \tag{5}$$

$$\mathbf{C}_i = \mathrm{BERT}(\mathcal{C}_i), \tag{6}$$

where $\mathbf{t}_i \in \mathbb{R}^{1\times d}$, $\mathbf{B}_i = \{\mathbf{b}_i^1, \mathbf{b}_i^2, \dots, \mathbf{b}_i^{|\mathcal{B}_i|}\} \in \mathbb{R}^{|\mathcal{B}_i|\times d}$, and $\mathbf{C}_i = \{\mathbf{c}_i^1, \mathbf{c}_i^2, \dots, \mathbf{c}_i^{|\mathcal{C}_i|}\} \in \mathbb{R}^{|\mathcal{C}_i|\times d}$[5] represent text vectors obtained for the original news, background information, and inconsistency points, respectively. d is the dimensional of vectors.

3.4. Semantic-aware feature alignment

Once the latent representations for the original news content and external knowledge are obtained, existing methods usually either concatenate them directly for prediction or fuse them without considering semantic alignment. However, despite the high-quality content generated by state-of-the-art LLMs, they still suffer from the phenomenon of hallucination (Zhang et al., 2023), where the generated content may fail to match user prompts, contradict previously generated context, or deviate from established world knowledge. This introduces **uncertainties** into our model — the generated external knowledge may not effectively assist in determining whether the original text is fake and could even compromise the model's detection accuracy to some extent.

To mitigate the effects of LLM's hallucinations and avoid injecting uncertainties into model prediction, we propose a **graph-based semantic-aware alignment** module (GSA). This module is responsible for aligning semantics between the original news article and the generated BIs and IPs in latent space.

3.4.1. Graph construction

We first create a fully connected semantic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{\{t\} \cup \mathcal{B} \cup \mathcal{C}\}$. Each node is associated with its learned text vector from the previous step, which serves as its node feature. As mentioned in Section 3.2, for different news items, the LLM will provide a various number of BIs and IPs, resulting in varying graph structures during their construction. Thus, we introduced two additional virtual nodes B and C, instead of using zero-padding to equalize the number of BIs and IPs across all news items. These two nodes are then connected to each node within $\mathcal{B}$ and $\mathcal{C}$, respectively. Meanwhile, the connections among t, $\{\mathbf{b}_n\}_{n=1}^{|\mathcal{B}|}$, and $\{\mathbf{c}_m\}_{m=1}^{|\mathcal{C}|}$ are removed, and new edges among t, B, and C are introduced. Fig. 2 part (b) shows a toy example of the constructed graph. By doing this: (1) Alignment computational costs are reduced, due to a smaller graph after edge reduction; (2) Semantic alignment effectiveness is improved. This improvement occurs because we first perform internal alignment within BIs and IPs, which enhances semantic consistency and reduces ambiguity among the items involved. The interaction between these items and their corresponding virtual semantic nodes, B and C, leads to more robust representations for BIs and IPs, respectively. Finally, external alignment is conducted between t, B, and C, alleviating the possibility of harmful information being injected into $\mathbf{t}$.

3.4.2. Feature alignment

After constructing the semantic graph, our model proceeds to the alignment step. We leverage Graph Attention Networks (GAT) (Veličković, Cucurull, Casanova, Romero, Liò, & Bengio, 2018) to assign unique importance scores to neighboring nodes with respect to the central node $\mathbf{t}$. Most of the nodes in $\mathcal{G}$ correspond to learned representations, so we directly use these representations as input node features to GAT. For the two virtual nodes, however, no specific representations are initially available. Therefore, various initialization methods can be utilized, such as all-zero-vector, random vector, or pooling based on BIs' or IPs' representations. We discuss the impact of these different settings on model performance in Section 4.6. In this work, we use $\mathbf{U}$ to denote the feature matrix of the nodes in $\mathcal{G}$, where the nodes are ordered as $\{t, B, C, b_1, \dots, b_{|\mathcal{B}|}, c_1, \dots, c_{|\mathcal{C}|}\}$. For simplicity, we represent the feature matrix as $\mathbf{U} = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$, where $n = 3 + |\mathcal{B}| + |\mathcal{C}|$, with $\mathbf{u}_1 = \mathbf{t}$, and so on.

In contrast to the original GAT, we integrate a multi-semantic fusion mechanism, which enhances GAT's ability to extract semantic information from complex interactions between nodes. Specifically, this mechanism is implemented by introducing multi-dimensional relationships into the edge feature computation between two nodes u_i and u_j, including the inter-node difference $|\mathbf{u}_i - \mathbf{u}_j|$ and the similarity $\mathbf{u}_i \cdot \mathbf{u}_j$. Notably, in our implementation, we use *multi-layer* (L), *multi-head* (K) settings in our multi-semantic aware GAT, where each layer l corresponds to K attention heads. For instance, at layer $l+1$, we have:

$$\begin{aligned} e_{ij}^{l+1,k} = \mathbf{a}_{l+1,k}^{\top} \Big(&\mathbf{W}_u^{l+1,k}\mathbf{h}_i^l + \mathbf{W}_u^{l+1,k}\mathbf{h}_j^l \\ &+ \mathbf{W}_d^{l+1,k}|\mathbf{h}_i^l - \mathbf{h}_j^l| + \mathbf{W}_s^{l+1,k}(\mathbf{h}_i^l \cdot \mathbf{h}_j^l)\Big), \end{aligned} \tag{7}$$

$$\alpha_{ij}^{l+1,k} = \frac{\exp\left(\mathrm{LeakyReLU}\left(e_{ij}^{l+1,k}\right)\right)}{\sum_{*\in\mathcal{N}(i)} \exp\left(\mathrm{LeakyReLU}\left(e_{i*}^{l+1,k}\right)\right)}, \tag{8}$$

where $l+1, k$ refers to the $l+1$-th layer, and the kth attention head. $\mathbf{W}_u \in \mathbb{R}^{d^{l+1}\times d^l}$ is a shared weight matrix, responsible for the linear transformation of the input node embeddings, including both the target node and its neighbors. $\mathbf{W}_d \in \mathbb{R}^{d^{l+1}\times 1}$ and $\mathbf{W}_s \in \mathbb{R}^{d^{l+1}\times 1}$ are two trainable weight matrices applied to the nodes' semantic differences and similarities, respectively. $\mathbf{a} \in \mathbb{R}^{4d^{l+1}}$ is a weight vector. LeakyReLU is used to ensure nonlinearity. Subsequently, the updated node representation is calculated based on its attention score with all neighbors, formulated as follows:

$$\mathbf{h}_i^{l+1,k} = \sigma\left(\mathbf{W}_u^{l+1,k}\mathbf{h}_i^l + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}^{l+1,k}\mathbf{W}_h^{l+1,k}\mathbf{h}_j^{l+1,k}\right), \tag{9}$$

where $\mathbf{h}_i^{l+1,k} \in \mathbb{R}^{d^{l+1}}$, and σ denotes the nonlinear activation function. Notably, the initial input node embedding for the first layer is $\mathbf{h}_i^0 = \mathbf{u}_i \in \mathbf{U}$. Subsequently, to obtain the final node representation at $l+1$ layer, we aggregate the outputs from different attention heads as:

$$\mathbf{h}_i^{l+1} = \Big\|_{k=1}^{K} \mathbf{h}_i^{l+1,k}. \tag{10}$$

Herein, $\|$ represents the aggregation function. After L layer processing, the final node representations matrix is denoted as $\mathbf{H} \in \mathbb{R}^{n\times d} = \{\mathbf{h}_1, \dots, \mathbf{h}_n\}$. Then we extract the node representations for node t, B, and C from $\mathbf{H}$ via a look-up operation for further processing. This practice is based on our assumption that, through the multi-semantic GAT layer, knowledge is transferred from external knowledge nodes into the two virtual nodes, while also ensuring that the semantics of these two nodes are aligned with node t.

3.5. Knowledge cross-distillation

Until now, the obtained representations can be directly used as features for the final prediction. However, in this approach, once the model is well-trained and deployed to infer unseen samples, each prediction will require the model to request external knowledge from the LLM. This leads to increased inference time and makes our model heavily reliant on the LLM service. To streamline our model and reduce reliance on LLMs in future deployments, we propose employing distillation techniques to train two robust distillation injectors for t, which can absorb sufficient predictive information from B and C, respectively. Specifically, we grounded in information bottleneck (IB) theory (Tishby et al., 2001), and design a Knowledge Cross-Distillation (KCD) module, which uses variational inference to reconstruct the objective of IB, facilitating knowledge exchange between t and B, as well as t and C during the training phase.

[5] We omit the subscript i in the following sections for simplicity.

3.5.1. KCD structure

The structure of KCD is shown in Fig. 2, where it is clear that KCD's inputs are the representations of t, B, and C from the previous semantic alignment layer, i.e., $\mathbf{h}_t$, $\mathbf{h}_B$ and $\mathbf{h}_C$. KCD consists of two distillation injectors ($\mathrm{E}_B(\cdot)$ and $\mathrm{E}_C(\cdot)$) and two parameter-shared information bottlenecks (IBs) (IB_B and IB_C). Each distillation encoder facilitates knowledge injection from B and C, respectively, while the IBs guide the distillation process based on the information bottleneck theory.

3.5.2. Distillation objective

Taking the knowledge distiDistillation Objectivellation between $\mathbf{h}_t$ and $\mathbf{h}_B$ as an example (vice versa for $\mathbf{h}_t$ and $\mathbf{h}_C$), we then illustrate the distillation principle of KCD through a detailed discussion based on IB theory, and further define its optimization object.

Now, we have two different predictors $\mathbf{h}_t$ and $\mathbf{h}_B$ on hand, each containing sufficient information about the target y, our goal is to train a distillation encoder E_B to capture the consistent information accessible from both $\mathbf{h}_t$ and $\mathbf{h}_B$, while preserving the necessary label information. To achieve this goal, we introduce a shared information bottleneck IB_B to collaborate with E_B in performing distillation by maximizing the amount of view-consistent information — the mutual information between the outputs of E_B and IB_B, i.e., $I(\mathbf{z}_t^B;\mathbf{v}_B)$ and $I(\mathbf{z}_B;\mathbf{v}_t^B)$. Here, $\mathbf{v}_t,\mathbf{v}_B = E_B(\mathbf{h}_t,\mathbf{h}_B)$ and $\mathbf{z}_t^B,\mathbf{z}_B = IB_B(\mathbf{v}_t,\mathbf{v}_B)$. In the following discussion, we use $I(\mathbf{z}_t^B;\mathbf{v}_B)$ as an example and first apply the chain rule to factorize it as follows (Federici, Dutta, Forré, Kushman, & Akata, 2020):

$$I(\mathbf{z}_t^B;\mathbf{v}_B) = \underbrace{I(\mathbf{v}_B;\mathbf{z}_t^B|y)}_{\text{superfluous}} + \underbrace{I(\mathbf{z}_t^B;y)}_{\text{predictive}}. \tag{11}$$

Herein, $I(\mathbf{v}_B;\mathbf{z}_t^B|y)$ denotes the irrelevant information encoded in $\mathbf{z}_t^B$ w.r.t. y, while $I(\mathbf{z}_t^B;y)$ represents the amount of information maintained in $\mathbf{z}_t^B$ that is relevant to y. Thus, we reformulate the maximization of $I(\mathbf{z}_t^B;\mathbf{v}_B)$ as the task of simultaneously maximizing $I(\mathbf{z}_t^B;y)$ and minimizing $I(\mathbf{v}_B;\mathbf{z}_t^B|y)$.

According to the data processing inequality (Burt, Ober, Garriga-Alonso, & van der Wilk, 2020), which gives $I(\mathbf{z}_t^B;y) \le I(\mathbf{v}_B;y)$, we can rephrase Eq. (11) as:

$$I(\mathbf{z}_t^B;\mathbf{v}_B) \le I(\mathbf{v}_B;\mathbf{z}_t^B|y) + I(\mathbf{v}_B;y). \tag{12}$$

Introducing $I(\mathbf{v}_t;y)$ decomposes the maximization of $I(\mathbf{z}_t^B;\mathbf{v}_B)$ into three sub-optimizations (Tian et al., 2021): maximizing $I(\mathbf{v}_B;y)$, minimizing $I(\mathbf{v}_B;y) - I(\mathbf{z}_t^B;y)$ and $I(\mathbf{v}_B;\mathbf{z}_t^B|y)$. Here, $I(\mathbf{v}_t;y)$ can be optimized through our supervised detection task, and minimizing $I(\mathbf{v}_B;y) - I(\mathbf{z}_t^B;y)$ forces $I(\mathbf{z}_t^B;y)$ to approximate $I(\mathbf{v}_B;y)$, which intuitively reduce $I(\mathbf{v}_B;\mathbf{z}_t^B|y)$. Therefore, the optimization can be simplified to:

$$\min I(\mathbf{v}_B;y) - I(\mathbf{z}_t^B;y). \tag{13}$$

Recalling the definition of mutual information (Belghazi et al., 2018), we can rewrite $I(\mathbf{v}_B;y) - I(\mathbf{z}_t^B;y)$ by introducing entropy as:

$$\begin{aligned} I(\mathbf{v}_B;y) - I(\mathbf{z}_t^B;y) &= H(y) - H(y|\mathbf{v}_B) - (H(y) - H(y|\mathbf{z}_t^B)) \\ &= H(y|\mathbf{z}_t^B) - H(y|\mathbf{v}_B). \end{aligned} \tag{14}$$

Thus, Eq. (13) is equal to:

$$\min H(y|\mathbf{z}_t^B) - H(y|\mathbf{v}_B). \tag{15}$$

To derive the upper bound of Eq. (15), we rephrase the conditional entropy in its internal form as:

$$\begin{aligned} & H(y|\mathbf{z}_t^B) - H(y|\mathbf{v}_B) = \\ & - \int p(\mathbf{z}_t^B) d_{\mathbf{z}_t^B} \int p(y|\mathbf{z}_t^B) \log p(y|\mathbf{z}_t^B) d_y \\ & + \int p(\mathbf{v}_B) d_{\mathbf{v}_B} \int p(y|\mathbf{v}_B) \log p(y|\mathbf{v}_B) d_y = \\ & - \iint p(\mathbf{z}_t^B) p(y|\mathbf{z}_t^B) \log \left[\frac{p(y|\mathbf{z}_t^B)}{p(y|\mathbf{v}_B)} p(y|\mathbf{v}_B) \right] d_{\mathbf{z}_t^B} d_y \end{aligned} \tag{16}$$

$$\begin{aligned} & + \iint p(\mathbf{v}_B) p(y|\mathbf{v}_B) \log \left[\frac{p(y|\mathbf{v}_B)}{p(y|\mathbf{z}_t^B)} p(y|\mathbf{z}_t^B) \right] d_{\mathbf{v}_B} d_y = \\ & - \iint p(\mathbf{z}_t^B) p(y|\mathbf{z}_t^B) \log \frac{p(y|\mathbf{z}_t^B)}{p(y|\mathbf{v}_B)} d_{\mathbf{z}_t^B} d_y \\ & - \iint p(\mathbf{z}_t^B) p(y|\mathbf{z}_t^B) \log p(y|\mathbf{v}_B) d_{\mathbf{z}_t^B} d_y \\ & + \iint p(\mathbf{v}_B) p(y|\mathbf{v}_B) \log \frac{p(y|\mathbf{v}_B)}{p(y|\mathbf{z}_t^B)} d_{\mathbf{v}_B} d_y \\ & + \iint p(\mathbf{v}_B) p(y|\mathbf{v}_B) \log p(y|\mathbf{z}_t^B) d_{\mathbf{v}_B} d_y = \\ & - \int p(\mathbf{z}_t^B) D_{KL} \left[p(y|\mathbf{z}_t^B) \parallel p(y|\mathbf{v}_B) \right] d_{\mathbf{z}_t^B} \\ & + \int p(\mathbf{v}_B) D_{KL} \left[p(y|\mathbf{v}_B) \parallel p(y|\mathbf{z}_t^B) \right] d_{\mathbf{v}_B} \\ & + \int p(y) \log \frac{p(y|\mathbf{z}_t^B)}{p(y|\mathbf{v}_B)} d_y \le \\ & \mathbb{E}_{\mathbf{v}_B \sim E_B(\theta_B)} \mathbb{E}_{\mathbf{z}_t^B \sim IB_B(\phi_B)} D_{KL} \left[p(y|\mathbf{v}_B) \parallel p(y|\mathbf{z}_t^B) \right] \\ & + \mathbb{E}_{\mathbf{v}_B \sim E_B(\theta_B)} \mathbb{E}_{\mathbf{z}_t^B \sim IB_B(\phi_B)} \log \frac{p(y|\mathbf{z}_t^B)}{p(y|\mathbf{v}_B)}, \end{aligned} \tag{17}$$

where θ_B and ϕ_B are parameters of E_B and IB_B, respectively. Subsequently, the distillation objective (i.e., Eq. (15)) is re-formalized as:

$$\begin{aligned} & \min_{\theta_B,\phi_B} \mathbb{E}_{\mathbf{v}_B \sim E_B(\theta_B)} \mathbb{E}_{\mathbf{z}_t^B \sim IB_B(\phi_B)} \\ & \left[D_{KL} \left[p(y|\mathbf{v}_B) \parallel p(y|\mathbf{z}_t^B) \right] + \log \frac{p(y|\mathbf{z}_t^B)}{p(y|\mathbf{v}_B)} \right]. \end{aligned} \tag{18}$$

Notably, minimizing the first term in Eq. (17) can explicitly approximate $p(y|\mathbf{z}_t^B)$ to $p(y|\mathbf{v}_B)$, thereby implicitly reducing the second term (Tian et al., 2021). Thus, we omit the second term in our implementation, that is,

$$\min_{\theta_B,\phi_B} \mathbb{E}_{\mathbf{v}_B \sim E_B(\theta_B)} \mathbb{E}_{\mathbf{z}_t^B \sim IB_B(\phi_B)} D_{KL} \left[p(y|\mathbf{v}_B) \parallel p(y|\mathbf{z}_t^B) \right]. \tag{19}$$

Afterward, we apply a similar derivation process to obtain the sub-objective for $I(\mathbf{z}_B;\mathbf{v}_t^B)$ and combine it with Eq. (18) to form the final objective for knowledge cross-distillation (KCD) between $\mathbf{h}_t$ and $\mathbf{h}_B$, i.e.,

$$\begin{aligned} \mathcal{L}_{KCD_B} = & \min_{\theta_B,\phi_B} \mathbb{E}_{\mathbf{v}_t^B,\mathbf{v}_B \sim E_B(\theta_B)} \mathbb{E}_{\mathbf{z}_t^B,\mathbf{z}_B \sim IB_B(\phi_B)} \\ & \left[D_{KL}[p(y|\mathbf{v}_B) \parallel p(y|\mathbf{z}_t^B)] + D_{KL}[p(y|\mathbf{v}_t^B) \parallel p(y|\mathbf{z}_B)] \right]. \end{aligned} \tag{20}$$

Similarly, the distillation objective between $\mathbf{h}_t$ and $\mathbf{h}_C$ is given as follows:

$$\begin{aligned} \mathcal{L}_{KCD_C} = & \min_{\theta_C,\phi_C} \mathbb{E}_{\mathbf{v}_t^C,\mathbf{v}_C \sim E_C(\theta)_C} \mathbb{E}_{\mathbf{z}_t^C,\mathbf{z}_C \sim IB_C(\phi_C)} \\ & \left[D_{KL}[p(y|\mathbf{v}_C) \parallel p(y|\mathbf{z}_t^C)] + D_{KL}[p(y|\mathbf{v}_t^C) \parallel p(y|\mathbf{z}_C)] \right]. \end{aligned} \tag{21}$$

3.6. Model training and inference

As shown in Fig. 2, the training and inference phases of LEKD rely on different components, which we will explain in detail in this section.

Training: As introduced in Section 3.2, we use the enriched dataset D_{aug} as training set. Each entry in D_{aug} comprises the original news article t, along with its corresponding background information $\mathcal{B}$ and inconsistency points C, both retrieved from LLM. During training, a learnable BERT model is initially employed to extract text vectors for input contents (Section 3.3). These vectors are then aligned within the same semantic space using the GSA module (Section 3.4). Subsequently, leveraging the information bottleneck theory, we train two distillation injectors to distill BIs and IPs information, respectively, ensuring the implicit accessibility of external knowledge during inference (Section 3.5). Finally, we concatenate the news vector $\mathbf{t}$ with its corresponding outputs from E_B and E_C, i.e., $\mathbf{v}_t^B$ and $\mathbf{v}_t^C$, for the final label prediction. Specifically, we fed the fused vector to a layer norm

Table 2
Dataset statistic.

Dataset	Weibo			GossipCop		
	Train	Val	Test	Train	Val	Test
Real	7,660	1,918	2,957	5,023	1,657	1,647
Fake	2,557	499	750	2,009	537	531
Period	2010–2017	2018	2018	2000–2017	2018	2018
Total	10,217	2,417	3,707	7,032	2,194	2,178

layer, followed by an MLP classifier. The final predicted label is denoted as $\hat{y}$. The formulation is defined as:

$$\hat{y} = \mathrm{MLP}(\mathrm{LN}(\mathbf{t} \oplus \mathbf{v}_t^B \oplus \mathbf{v}_t^C)). \tag{22}$$

The detection task is optimized by minimizing the cross-entropy between the predicted and ground truth labels:

$$\mathcal{L}_{CE} = -\left(y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})\right). \tag{23}$$

Recalling the two distillation objectives (ref. Eqs. (20) and (21)), the total loss function is given by:

$$\mathcal{L} = \alpha \mathcal{L}_{CE} + \beta * \mathcal{L}_{KCD_B} + \gamma * \mathcal{L}_{KCD_C}, \tag{24}$$

where α, β and γ are the trade-off hyperparameters.

Inference: During inference, only the news articles are used as input. For instance, a news article t is first embedded using the frozen BERT to obtain the text vector $\mathbf{t}$. $\mathbf{t}$ is then fed into the distillation injectors to obtain two external-knowledge-enhanced representations, i.e., $\mathbf{v}_t^B$ and $\mathbf{v}_t^C$. The final label is predicted based on the concatenation vector derived from $\mathbf{t}$, $\mathbf{v}_t^B$, and $\mathbf{v}_t^C$.

4. Evaluations

In this section, we discuss the effectiveness of our approach in detail. We start by outlining the datasets and experimental configurations used in our experiments, followed by a comprehensive analysis of our method's performance on these datasets.

4.1. Experimental settings

Dataset. We introduce two datasets in different languages for model evaluation: *Weibo* (Sheng et al., 2022) in Chinese and *GossipCop* (Shu, Mahudeswaran, Wang, Lee, & Liu, 2020) in English. The Weibo dataset contains a set of news published between 2010 and 2018, while the GossipCop dataset spans the period from 2000 to 2018. Following existing works (Hu, Sheng, et al., 2024; Zhu et al., 2022), we preprocess the datasets by applying deduplication and temporal data splitting, ensuring the prevention of data leakage, while simulating real-world temporal scenarios. The strategy for temporal data splitting involves first reordering the news articles by their publication time. Then, 40% of the most recently published news articles are randomly assigned to the test and validation sets in a 2:1 ratio, while the remaining 60% are used as the training set. Table 2 shows the statistical results for the dataset.

Baselines. We select the following representative content-centric detection methods as baselines:

- **LLM**. We directly fed the news articles to the LLM, asking it to determine the truthfulness of the given news. Notably, the prompts provided to the LLM contained no examples. In practice, we utilize GPT-3.5 Turbo. The detailed prompt settings used in our experiments for the LLM-based detector are outlined in Table 3.
- **BiGRU** (Cho, van Merriënboer, Bahdanau, & Bengio, 2014). A widely used backbone for text modeling. Specifically, we employ a single-layer BiGRU with a hidden size of 768, along with a masked attention layer to condense all hidden states into a news representation, which is then fed into an MLP for prediction.
- **EANN** (Wang et al., 2018). An adversarial neural network for multi-modal fake news detection. In our work, we use its text-only variant for comparison (i.e., only use Text-CNN to model news content). And the labels for the auxiliary event classification task are derived from clustering based on publication year.
- **BERT** (Devlin et al., 2019). The most popular pre-training model in NLP field. We use BERT to encode the tokens of the news content, and the extracted average embedding values are then fed into an MLP to produce the final prediction. The selection of BERT for the different datasets is consistent with our LEKD.
- **Emo** (Zhang et al., 2021). A BERT-based fake news detection model that introduces sentiment features into consideration. For a fair comparison, we use a simplified version that considers only the sentiment of the news content, excluding the sentiment from the comments
- **ENDEF** (Zhu et al., 2022). The module aims to remove entity bias through causal learning and can be easily integrated into existing content-centric detection models. In our experiments, we use BERT as the backbone, consistent with the choice of BERT used in our proposed model LEKD.
- **ARGD** (Hu, Sheng, et al., 2024). The latest work enhances SLMs by utilizing explanations extracted from LLMs, allowing the small models to better analyze news content. The goal of ARGD is to empower small detectors with the ability to flexibly choose useful rationales, which serve as references for the final judgment. In particular, we trained ARGD on the same training set as LEKD.

Evaluation Metrics. Besides the commonly used metrics, such as macro F1 score (macF1) and accuracy (ACC), we also follow previous works (Hu, Sheng, et al., 2024; Zhu et al., 2022) by reporting Area Under ROC (AUC), standardized partial AUC (spAUC), and the F1-score for each label ($\mathrm{F1}_{real}$ and $\mathrm{F1}_{fake}$) in our results. Here, we require the false positive rate (FPR) to be less than 10% according to Zhu et al. (2022). The calculation fo spAUC is defined as follows (McClish, 1989):

$$\mathrm{spAUC}_{\mathrm{FPR} \le \mathrm{maxfpr}} = \frac{1}{2}\left(1 + \frac{\mathrm{AUC}_{\mathrm{FPR} \le \mathrm{maxfpr}} - \mathrm{minarea}}{\mathrm{maxarea} - \mathrm{minarea}}\right),$$
$$\text{where } \mathrm{maxarea} = \mathrm{maxfpr},$$
$$\mathrm{minarea} = \frac{1}{2} \times \mathrm{maxfpr}, \tag{25}$$

Implementation Details. All baselines, including our LEKD are implemented using the PyTorch library, and accelerated by one NVIDIA RTX 4090 GPU. We utilize Adam optimizer (Kingma & Ba, 2014) with an initial learning rate set to 0.0001. Early stopping is adopted during training with a step size of 5. All methods are initialized with the same random seed, and use a batch size of 32. The maximum sequence length of the news article is uniformly set to 170. The other hyperparameters for each baseline follow the same settings as in the work by Zhu et al. (2022). For those methods that employ BERT as the backbone, including LEKD, we use *bert-base-chinese*[6] for the Weibo dataset and *bert-base-uncased*[7] for the GossipCop dataset, respectively.

In addition, for LEKD, we fixed the number of GAT layers and attention heads in the GSA to 3 and 32, respectively, for both the Weibo and GossipCop datasets. The initialization methods are set to Sum-pooling for Weibo and Min-pooling for GossipCop. The trade-off hyperparameters β and γ in the loss function are 600 and 200 for Weibo, and 200 and 600 for GossipCop, with α fixed at 1000 for both datasets.

4.2. Overall performance

Table 4 presents the performance of LEKD compared to the selected baselines on the two datasets. We have the following observations: (*O1*) Overall, we clearly observe that the proposed LEKD outperforms all

[6] https://huggingface.co/google-bert/bert-base-chinese
[7] https://huggingface.co/google-bert/bert-base-uncased

Table 3
Prompt Templates for the LLM-based detector.

English prompt templates	Chinese prompt templates
You are asked to decide whether this news article is true or false, and you can only answer either True or False. Here is the news: [t_i]	请你判断一下这篇新闻的真假,你只能回答True或者False这两个单词以下是新闻内容: [t_i]

Table 4
Performance comparison on two datasets. Bold indicates the best performance, while underline is the second best.

Method	Weibo						GossipCop					
	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC
LLM	0.634	0.767	0.501	0.683	0.721	0.533	0.612	0.808	0.417	0.711	0.610	0.531
BiGRU	0.687	0.882	0.492	0.808	0.799	0.620	0.776	0.894	0.658	0.839	0.861	0.735
EANN	0.716	0.892	0.541	0.825	0.813	0.659	0.788	0.901	0.675	0.849	0.872	0.750
BERT	0.749	0.901	0.596	0.842	0.860	0.691	0.780	0.892	0.668	0.837	0.853	0.740
ENDEF	0.766	0.904	0.627	0.848	0.879	0.716	0.785	0.895	0.675	0.842	0.872	0.742
Emo	0.763	0.907	0.619	0.851	0.869	0.714	0.775	0.898	0.652	0.843	0.866	0.739
ARGD	0.765	0.913	0.617	0.858	0.870	0.717	0.793	0.905	0.680	0.853	0.877	0.760
LEKD	**0.783**	**0.916**	**0.651**	**0.864**	**0.882**	**0.744**	**0.805**	**0.909**	**0.700**	**0.861**	**0.886**	**0.762**

baselines across all datasets in terms of every metric, demonstrating both the robustness and efficiency of LEKD. (*O2*) Among all baselines, LLM performs the worst, suggesting that while LLMs excel in text generation and comprehension, they still struggle to fully grasp human intentions behind the human-generated content and assess its credibility. The performance of BiGRU surpasses that of the LLM, owing to its specific training for fake news detection in a supervised manner. However, its limitations in handling long word sequences hinder further improvement. based text encoder, achieving better performance than the naive BiGRU. However, its performance still lags behind the BERT-based methods. (*O3*) The performances of the BERT-based baselines—BERT, ENDEF, Emo, and ARGD—are relatively close. We attribute this to the use of the same BERT backbone, resulting in similar representation learning abilities. Additionally, since each of ENDEF, Emo, and ARGD integrates novel modules to address different challenges in fake news detection, their performances are accordingly higher than naive BERT. (*O4*) Furthermore, for most of the metrics, ARGD surpasses other baselines. This is because ARGD is trained on the augmented dataset, enabling it to leverage external knowledge to enhance the model's feature extraction capabilities. (*O5*) Comparing the F1-scores across different labels, we observe a significant margin between the detection of real news and fake news. We argue that the reason lies in the skewed training sets of both datasets, where the real:fake ratio is approximately 3:1 and 2.5:1 for Weibo and GossipCop, respectively. The label imbalance leads all methods to tend to distinguish real news from fake news more accurately, but the reverse does not hold true. However, according to the spAUC results, our proposed LEKD misclassifies fewer fake news instances as real compared to other methods. To further prove this claim, we visualized the classification distribution for fake news on the Weibo and GossipCop test sets for both ARGD and our LEKD in Figs. 3 and 4. From where, we observe that, compared to ARGD, LEKD can more accurately detect most of the fake news, as indicated by the prediction scores in the range of 0.5 to 1. Additionally, LEKD achieves average scores of 0.62 and 0.649 on the Weibo and GossipCop datasets, respectively, both higher than ARGD. We counted the number of fake news articles misclassified by both ARGD and LEKD (see Fig. 5). These numbers imply that some fake news articles remain difficult to detect. However, from another perspective, it is clear that LEKDhas already significantly reduced this number to some extent. (*O6*) The performance gap between different datasets reveals that understanding Chinese remains more challenging than English.

Comparison of computational cost: We also statistic the computational cost for baselines and LEKD on two datasets, including training and inference time cost. The results are shown in Table 5. We find that both ARGD and LEKD incur higher training time costs compared to other news content-only methods, primarily due to the integration of external knowledge. However, during inference, LEKD significantly outperforms ARGD in terms of efficiency, achieving costs comparable to other simpler baselines (from BiGRU to Emo). This ability to integrate additional knowledge during training to enhance detection capabilities while maintaining low inference costs underscores the practicality and scalability of LEKD for real-world deployment.

Table 5
Time cost (s/epoch) for training (Inference).

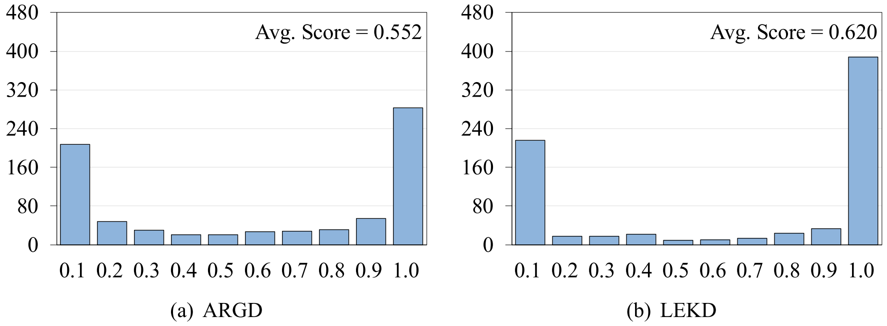
Model	Weibo	GossipCop
BiGRU	3.384(0.383)	2.393(0.399)
EANN	1.032(0.235)	0.799(0.201)
BERT	11.069(2.268)	7.761(2.223)
ENDEF	14.828(2.965)	10.366(2.950)
Emo	14.333(2.534)	10.009(2.546)
ARGD	85.084(9.261)	67.693(9.519)
LEKD	74.492(3.082)	63.846(3.460)

Comparison of LEKD and ARDG: Our LEKD, compared to ARGD–the latest work that employs LLMs as advisors to assist SLMs in fake news detection–shows the following key differences: *1. External knowledge generation*: LEKD employs a simpler zero-shot prompt, primarily leveraging the summarization ability and rich common sense inherent in LLMs, whereas ARGD explores more complex prompt techniques to generate both rationales and news labels, which are influenced by the reasoning capabilities of LLMs. *2. The use of external knowledge*: ARGD directly models the interaction between rationales and the original news text, while LEKD introduces a feature alignment module to mitigate contradictions between the original news content and the external knowledge. *3. Distillation strategies*: ARGD first trains a teacher model and then transfers rationale knowledge to a lightweight student model using a basic distillation strategy, which requires additional training. In contrast, LEKD requires only a single training phase; once trained, the corresponding modules can be invoked to implicitly generate latent external knowledge.

4.3. Ablation studies

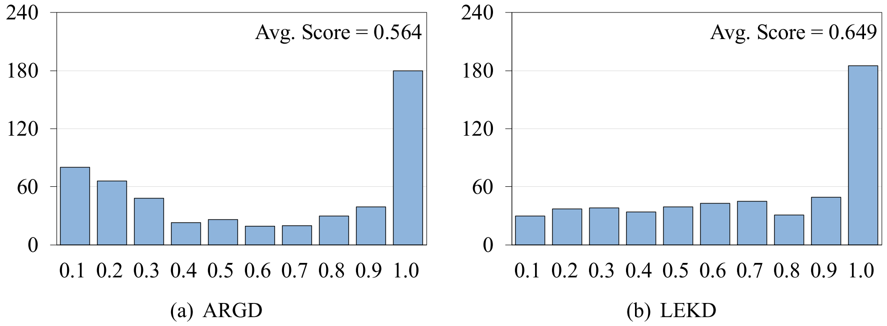
Herein, we perform ablation studies to evaluate the contribution of core components in LEKD. Specifically, we derive the following variants:

- $\text{LEKD}_{\mathbf{w/oGSA}}$ that removes the graph-based semantic-aware alignment module, and we directly aggregate the representations of BIs and IPs to form the text vectors for the two virtual nodes.



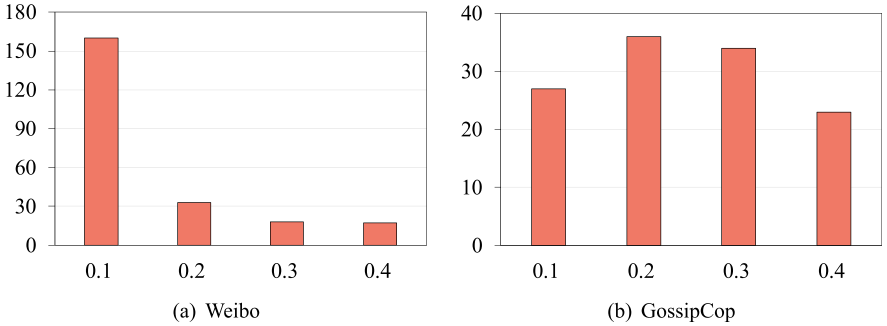
(a) ARGD (b) LEKD

Fig. 3. The distribution of prediction scores for fake news on the *Weibo* test sets for both ARGD and LEKD.



(a) ARGD (b) LEKD

Fig. 4. The distribution of prediction scores for fake news on the *GossipCop* test sets for both ARGD and LEKD.



(a) Weibo (b) GossipCop

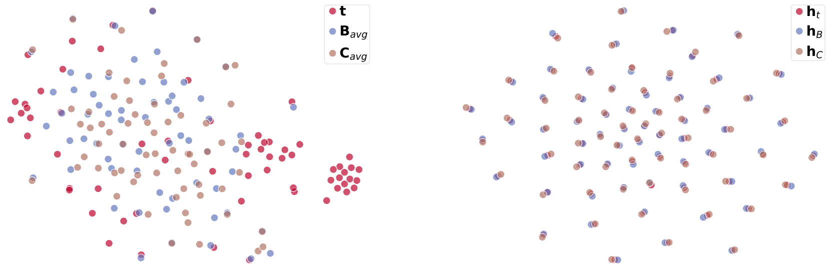
Fig. 5. The number of fake news articles incorrectly predicted by both LEKD and ARGD.

Table 6
Ablation study of LEKD.

Variant	Weibo						GossipCop					
	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC
$\mathrm{LEKD}_{\mathbf{w/o\ GSA}}$	0.767	0.906	0.629	0.850	0.891	0.717	0.790	0.897	0.684	0.845	0.879	0.759
$\mathrm{LEKD}_{\mathbf{w/o\ BIs}}$	0.778	0.914	0.643	0.861	0.881	0.732	0.796	0.904	0.688	0.853	0.887	0.764
$\mathrm{LEKD}_{\mathbf{w/o\ IPs}}$	0.769	0.915	0.623	0.862	0.882	0.729	0.785	0.903	0.667	0.850	0.869	0.747
$\mathrm{LEKD}_{\mathbf{w\ EXT}}$	0.767	0.904	0.630	0.848	0.852	0.707	0.797	0.901	0.693	0.850	0.778	0.709
LEKD	**0.783**	**0.916**	**0.651**	**0.864**	**0.882**	**0.744**	**0.805**	**0.909**	**0.700**	**0.861**	**0.886**	**0.762**

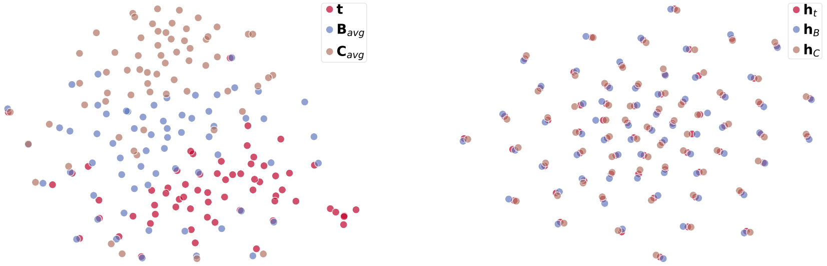
- $\mathrm{LEKD}_{\mathbf{w/oBIs}}$, which retains only Inconsistent points as external knowledge.
- $\mathrm{LEKD}_{\mathbf{w/oIPs}}$ uses only background information as external knowledge for model training.
- $\mathrm{LEKD}_{\mathbf{wEXT}}$ removes the Knowledge Cross-Distillation module and directly utilizes all external knowledge during the inference stage.

The experiments are reported in Table 6, and we can draw the following conclusion: the removal of GSA ($\mathrm{LEKD}_{\mathbf{w/oGSA}}$) resulted in a larger performance decrease compared to the other two variants, indicating that GSA significantly improves the semantic accuracy of external knowledge with respect to the original news article. Additionally, these results highlight the importance of semantic alignment.



(a) Before (b) After

Fig. 6. Visualization of the representations of selected Weibo test news articles before and after applying GSA, using t-SNE.



(a) Before (b) After

Fig. 7. Visualization of the representations of selected GossipCop test news articles before and after applying GSA, using t-SNE.

Comparing $\text{LEKD}_{\mathbf{w/oBIs}}$ to $\text{LEKD}_{\mathbf{w/oIPs}}$ shows that IPs are relatively more informative than BIs, as $\text{LEKD}_{\mathbf{w/oBIs}} \geq \text{LEKD}_{\mathbf{w/oIPs}}$. Naturally, fully incorporating both into model training results in the best performance. Furthermore, even when external knowledge was incorporated during the testing stage (cf. $\text{LEKD}_{\mathbf{wEXT}}$), the model's performance did not improve but instead decreased. This outcome highlights the effectiveness of the KCD design, as it significantly reduces redundancy while retaining only the most relevant information.

In the following, we investigate the effectiveness of the proposed GSA module in achieving feature alignment. Specifically, we first randomly select 60 news articles from the test sets of the Weibo and GossipCop datasets, respectively. Then, we use *t*-SNE (Van der Maaten & Hinton, 2008) to visualize the representations of the original news text and external knowledge, both before and after applying the GSA. These include the text embeddings from the BERT Encoder and the outputs from the GSA. For the text embeddings, we visualize $\mathbf{t}$, along with the averaged knowledge embeddings for both BIs and IPs, i.e., $\mathbf{B}_{avg} = \text{mean}(\mathbf{B})$ and $\mathbf{C}_{avg} = \text{mean}(\mathbf{C})$. For the GSA outputs, we visualize $\mathbf{h}_t$, $\mathbf{h}_B$, and $\mathbf{h}_C$.

The visualization results are shown in Fig. 6. The first two figures (ref. Figs. 6(a) and 7(a)) depict the latent representations before applying GSA on selected news articles, while the last two (Figs. 6(b) and 7(b)) display the results after feature alignment. It is evident that, before applying GSA, there is a significant semantic gap between the original text and its corresponding external knowledge. However, after feature alignment, on the one hand, the GSA effectively transfers knowledge from BIs and IPs to their corresponding virtual nodes. On the other hand, it realizes that the original text representations and the representations of external knowledge virtual nodes are well-aligned within the same semantic space.

4.4. Generalization and distillation effectiveness analysis

In this section, we first perform a generalization test on two new supplementary datasets using models pre-trained on the Weibo and GossipCop datasets, thereby demonstrating the competitive generalization ability of our LEKD. The statistics of the supplementary datasets are listed in Table 7. The datasets are collected from Weibo21 (Nan, Cao, Zhu, Wang, & Li, 2021) and GossipCop++ (Su, Cardie, & Nakov, 2024). Specifically, for the Weibo21 dataset, we adopt the test set in our experiments. For GossipCop++, we collect the MF and MR subsets of the entire dataset, where MF denotes machine-generated fake news and MR represents machine-generated real news. Notably, these news articles are generated by mimicking the style and topics of the original GossipCop dataset.

Table 7
Statistic of supplementary datasets.

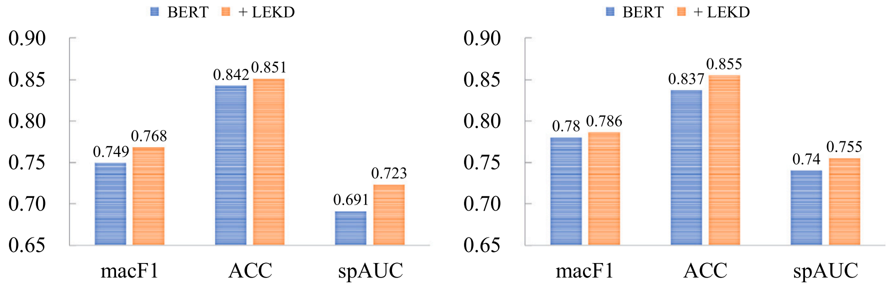
Dataset	Real	Fake	Total
Weibo21	1,137	814	1,951
Gossipcop++	4,169	4,084	8,253

The results are presented in Table 8. It is evident that our LEKD outperforms the selected baseline methods, i.e., BERT and ARGD, with significant improvement. We owe this to our method's ability to implicitly generate external knowledge as a supplement to the original news, enabling the model to capture different perspectives of the news and make more accurate predictions. Similar to LEKD, ARGD uses a basic distillation strategy to transfer external knowledge from a well-trained teacher model to a lightweight student model, allowing it to surpass naive BERT. However, ARGD's ability to learn external knowledge still falls behind ours. In addition, our LEKD does not require training a student model. Furthermore, the results on the GossipCop++ dataset show remarkable advances compared to the Weibo21 dataset. The reason is that the news articles in GossipCop++ cover the same topic domain and share a similar writing style to the original GossipCop dataset, which we used to train our model.

As we have stated in Section 3.5, our LEKD can implicitly generate latent external knowledge during the inference phase. To validate this claim, we conduct experiments by integrating the two pre-trained

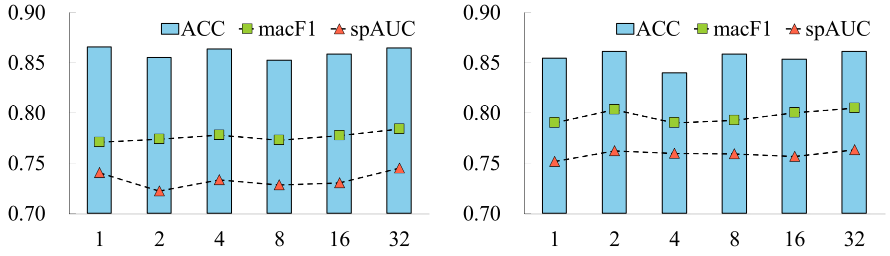
Table 8
Generalization test on Weibo21 and GossipCop++ datasets.

Method	Weibo21						GossipCop++					
	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC
BERT	0.721	0.805	0.636	0.746	0.716	0.614	0.756	0.802	0.710	0.765	0.763	0.695
ARGD	0.766	0.839	0.693	0.789	0.758	0.693	0.779	0.814	0.743	0.784	0.783	0.701
LEKD	**0.772**	**0.842**	**0.702**	**0.793**	**0.764**	**0.698**	**0.806**	**0.833**	**0.780**	**0.810**	**0.808**	**0.720**



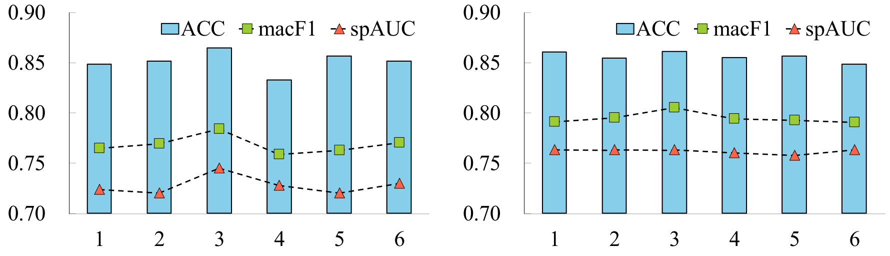
(a) Weibo (b) GossipCop

Fig. 8. Distillation effectiveness test of the injectors E_B and E_C. **BERT** refers to the basic BERT model trained on the original dataset, while **+LEKD** represents the BERT model enhanced with latent external knowledge, generated by the frozen injectors E_B and E_C during training.



(a) Weibo (b) GossipCop

Fig. 9. The impact of the number of attention heads K in the GSA module.



(a) Weibo (b) GossipCop

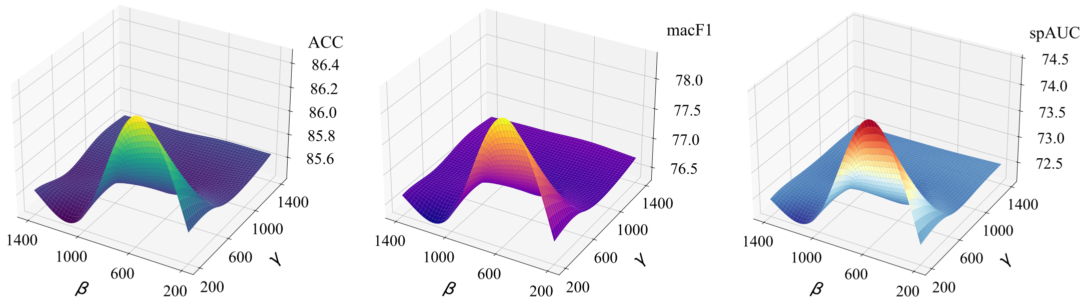
Fig. 10. The impact of the number of GAT layers L in the GSA module.

knowledge distillation injectors, i.e., E_B and E_C, into the basic BERT model. Specifically, we train a basic BERT model by concatenating the latent features for background information and inconsistency points, generated by the frozen E_B and E_C, respectively, with the original news text embedding to serve as detection cues for an MLP detector. Fig. 8 displays the results on two datasets. We observe that after incorporating the external latent knowledge generated by LEKD, the performance of the basic BERT model significantly improves compared to its standard version. Notably, the increase in macF1 and spAUC values highlights the model's enhanced ability to handle label-imbalance scenarios, allowing it to more accurately distinguish fake news from real news. In sum, the performance improvement of the BERT model strongly supports our claim that the KCD module effectively distills external knowledge into E_B and E_C.

4.5. Sensitivity analysis of hyperparameters

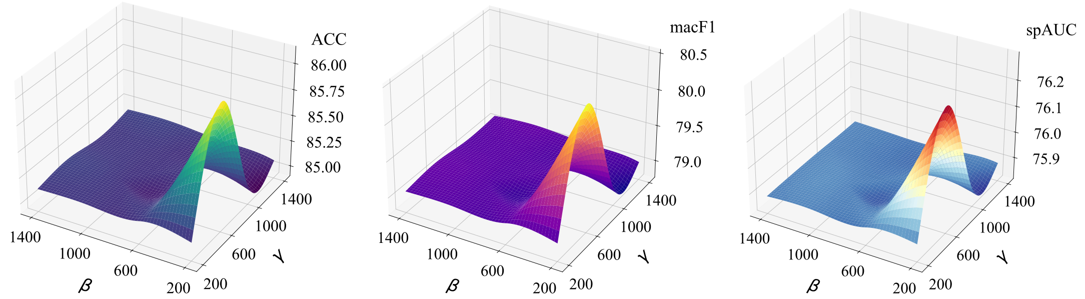
In this section, we analyze the impact of core hyperparameters that could potentially influence the LEKD's detection performance: (1) the number of attention heads K, (2) the number of GAT layers L in GSA, and (3) the loss balance hyperparameter β and γ. The experiments are conducted on both the Weibo and GossipCop datasets, and results are presented using the macF1, ACC, and spAUC metrics. Notably, in practice, we employ a One-Factor-At-A-Time (OFAT) approach, where we change only one hyperparameter at a time while keeping all other hyperparameters fixed at their optimal values, as discussed in Section 4.1.

Figs. 9 and 10 illustrate the model's performance with different numbers of attention heads per GAT layer and various total numbers of GAT layers, respectively. We observe that the optimal settings in the



(a) ACC (b) macF1 (c) spAUC

Fig. 11. Performance changes with the adjustment of the tradeoff hyperparameters β and γ on the Weibo dataset.



(a) ACC (b) macF1 (c) spAUC

Fig. 12. Performance changes with the adjustment of the tradeoff hyperparameters β and γ on the GossipCop dataset.

GSA module for both the Weibo and GossipCop datasets are $K = 32$ and $L = 3$, suggesting that fewer layers with relatively more attention heads result in better performance. However, this conclusion is specific to the Weibo and GossipCop datasets.

In Figs. 11 and 12, we plot the performance of LEKD on the Weibo and GossipCop datasets with varying values of β and γ in the range of [200 − 1400]. We note that, in most cases, smaller values tend to yield better performance. Specifically, $\beta = 600$ and $\gamma = 200$ are the optimal settings for Weibo, while $\beta = 200$ and $\gamma = 600$ are optimal for GossipCop. However, the choice of β and γ appears to be independent, as the experimental results show no evidence of a relationship between them. Additionally, we can infer that background information and inconsistency points hold different levels of importance in improving detection accuracy across the two datasets. Notably, the hyperparameter value is amplified by 100 times here to better observe the changes in the sub-loss components, as the original values of $\mathcal{L}_{KCD_B}$ and $\mathcal{L}_{KCD_C}$ are extremely small compared to $\mathcal{L}_{CE}$. In practice, the weight of $\mathcal{L}_{CE}$ is fixed at $\alpha = 1000$.

4.6. Discussion on initial embedding of virtual node

As we mentioned in Section 3.4, the method for initializing the embeddings of the two virtual nodes B and C is not fixed. Thus, in this section, we conduct experiments by utilizing different initialization methods. Specifically, we use (1) max-pooling (max) based on BIs and IPs, i.e., $\mathsf{B} = \max(\mathbf{b}^1, \cdot, \mathbf{b}^{|\mathcal{B}|})$ and $\mathsf{C} = \max(\mathbf{c}^1, \cdot, \mathbf{c}^{|\mathcal{C}|})$. Similarly, we can apply (2) min-pooling (min), (3) sum-pooling (sum), and (4) mean-pooling (mean). (5) We also select the most relevant elements to $\mathbf{t}$ from $\mathbf{B}$ and $\mathbf{C}$ as the initial embeddings for B and C, respectively, by calculating the cosine similarities between each item in $\mathbf{B}$ and $\mathbf{C}$ and $\mathbf{t}$ (sim). Additionally, we employ (6) all-zero vectors (zero) and (7) randomly generated vectors (random) as their initial embeddings.

Fig. 13 depicts the performance with different initialization methods. We see that different choices result in varying performance on different datasets. Herein, 'random' and 'min' are the optimal choices for the Weibo and GossipCop datasets, respectively. On both datasets, the 'sim' and 'zero' initialization methods perform slightly worse than the others. These unstable results reveal that there is no strict rule for selecting the initialization methods. However, the results of the 'random' initialization perform relatively well on both datasets, suggesting that the 'random' method is a strong choice for the default setting.

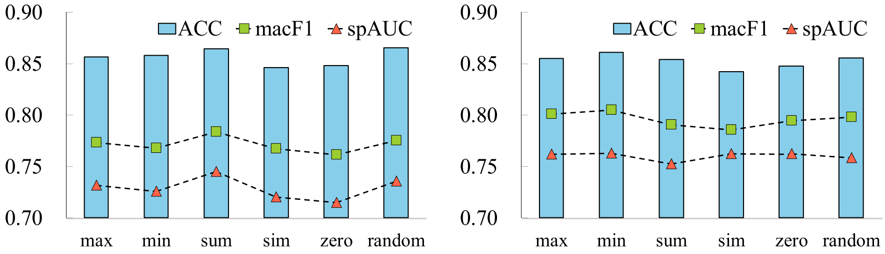
4.7. Comparison of different LLMs

To assess whether the model's performance is influenced by the selection of LLMs, we conducted experiments using four different LLMs, comprising two closed-source models and two open-source models:

- GPT: The primary choice of our experiments, utilizing the gpt-3.5-turbo-0125 version.
- ERNIE[8]: A knowledge-enhanced model by Baidu, we select the ERNIE-3.5-128K version in the experiment.
- Yi[9]: An open-source bilingual language model developed by 01.AI, we use the Yi-34B-Chat version[10] for comparison.
- Llama[11]: An open-source LLM developed by Meta. In this experiment, we employ the Llama-3.1-8B-Instruct version.[12]

Notably, all LLMs use the same prompt template to acquire external knowledge, as illustrated in Table 1. The configuration of each LLM used in our experiments is outlined in Table 9. Due to GPU limitations, GPT, ERNIE, and YI were accessed via their official APIs. For LLAMA, we selected the 8B-parameter version, which allowed generation on our local GPU. The corresponding generation times for external knowledge per news item are also provided. The experiments are conducted on

[8] https://github.com/PaddlePaddle/ERNIE
[9] https://github.com/01-ai/Yi
[10] https://huggingface.co/01-ai/Yi-34B-Chat
[11] https://www.llama.com/
[12] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct



(a) Weibo

(b) GossipCop

Fig. 13. Different initialization methods for virtual nodes.

Table 9
Technical details of LLMs.

LLM	Version	Temperature	top_p	Invocation Method	Avg. gen time(s/Per news)
GPT	GPT-3.5-turbo-0125	1	1	API	6.132
ERNIE	ERNIE-3.5-128K	0.8	0.8	API	7.627
YI	Yi-34B-Chat	0.6	0.8	API	9.606
LLAMA	Llama-3.1-8B-Instruct	0.6	0.9	Local	6.355

Table 10
Comparison of Different LLMs in LEKD.

Variant	Weibo						GossipCop					
	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC	macF1	$F1_{real}$	$F1_{fake}$	ACC	AUC	spAUC
LEKD$_{GPT}$	0.783	**0.916**	0.651	**0.864**	0.882	0.744	**0.805**	**0.909**	**0.700**	**0.861**	**0.886**	**0.762**
LEKD$_{ERNIE}$	0.783	0.913	0.652	0.861	0.881	0.738	0.796	0.905	0.687	0.854	0.875	0.759
LEKD$_{YI}$	**0.786**	0.910	**0.661**	0.858	**0.891**	**0.748**	0.793	0.903	0.684	0.851	0.876	0.759
LEKD$_{LLAMA}$	0.780	0.910	0.650	0.857	0.885	0.747	0.797	0.906	0.688	0.855	0.876	0.761

two datasets, and Table 10 presents the experimental results. Overall, the performance differences among the various LLMs are not significant, with GPT demonstrating relatively superior performance across all datasets. Therefore, we suggest that there is no specific rule for selecting LLMs, and the choice should be guided by experimental results and the specific requirements of the task.

4.8. Discussion on hallucination — A case study

In this section, we present a case study to illustrate the hallucination phenomenon in LLM-generated content, including our proposed approach, which leverages LLMs to generate external knowledge, and the results of directly employing LLMs as fact-checkers. The case study is detailed in Table 11, with hallucinated information highlighted using a red background. As shown in Table 11, LLMs exhibit fewer hallucinations when generating simple, concise knowledge from clear, well-defined prompts compared to their use in complex reasoning tasks. Specifically, in this case study, the background information and consistency-related content generated by the LLM align accurately with the input. However, the final line, in which the LLM is employed as a fact-checker, clearly suffers from a faithfulness hallucination (Huang et al., 2025). The response contradicts the timeline provided in the original article, where Nikki Bella and John Cena's wedding planned for May 5 was canceled due to their breakup on April 15. In contrast, the LLM incorrectly asserts that the couple would still marry on May 15 despite their breakup. This clearly highlights the increased risk of hallucination associated with complex reasoning tasks.

5. Conclusion

In this work, we propose a novel text-centric fake news detector, LEKD, which builds upon the foundations of current state-of-the-art paradigms, including SLM-based, external-knowledge-enhanced, and LLM-based methods, seamlessly integrating their strengths into a unified model. Specifically, LEKD leverages an LLM to provide complementary information, such as background information and inconsistency points based on the news text content. Notably, this knowledge acquisition step is conducted only on the training set. Next, LEKD uses a pre-trained SLM to generate textual representations for both the original news text and the external knowledge. Subsequently, a graph-based semantic-aware feature alignment module is introduced to mitigate hallucinations and ambiguity in external knowledge compared to the original news text. To ensure that latent external knowledge can be implicitly generated during the inference stage, LEKD also incorporates an information bottleneck-based knowledge cross-distillation component. In conclusion, we conducted extensive experiments on two real-world datasets, which demonstrated that LEKD outperforms existing state-of-the-art text-centric baselines, and that each component of LEKD is indispensable.

As future work, we will develop methods specifically designed to combat LLM-generated misinformation. Additionally, we plan to extend our model to a multimodal scenario. On the one hand, we plan to explore large vision-language models for detecting fake news and generating complementary information for both textual and visual

Table 11
Case study of hallucinations in LLM-generated Content (Hallucinated content highlighted in Red).

News article	Before Nikki Bella and John Cena announced on Sunday, April 15, that they had called it quits, they spent six years raving about each other. After being friends for years, the WWE Superstars began dating in 2012. Five years later, he proposed during the live broadcast of Wrestlemania 33 in April 2017, telling the world that he had been waiting for a very long time to ask for her hand in marriage . Cena, 40, and Bella, 34, were set to tie the knot on May 5 in Mexico. On Sunday, Us Weekly was the first to report that the pair had ended their engagement. " While this decision was a difficult one, we continue to have a great deal of love and respect for one another," the former couple said. "We ask that you respect our privacy during this time in our lives." Us Weekly exclusively revealed on Thursday, May 31, that the couple have reconciled Scroll through the gallery to relive some of their best quotes about each other.
Background Information	•Nikki Bella and John Cena announced their breakup on April 15 after dating for six years and being engaged for a year. •Us Weekly reported on May 31 that the couple has reconciled.
Inconsistency Points	•The abrupt breakup after several years of dating and being engaged. •The quick reconciliation after the public breakup.
News Label	True
LLM as Fack-checker	Based on the information provided in the news article, it seems that the news is false. The article mentions that Nikki Bella and John Cena announced their breakup on Sunday, April 15, and were set to tie the knot on May 5 in Mexico. However, it also states that they reconciled on Thursday, May 31. This creates a contradiction in the timeline of events. Therefore, the news article appears to be **false.**

features. On the other hand, we will investigate effective retrieval-augmented generation methods for multimodal data, thereby enhancing embedding-based multimodal fake news detectors. Moreover, we plan to incorporate additional contextual factors, such as news diffusion patterns and user interactions, into our framework to enhance its adaptability and robustness in real-world scenarios. Given that news content often originates from diverse domains, extending our model to effectively handle cross-domain scenarios is a worthwhile research direction. Lastly, interpretability is crucial for detection models (Li et al., 2024), making it another important focus for our future work.

CRediT authorship contribution statement

Xueqin Chen: Writing – original draft, Methodology, Conceptualization. **Xiaoyu Huang:** Visualization, Validation, Software, Investigation. **Qiang Gao:** Writing – original draft, Validation, Resources, Methodology, Funding acquisition. **Li Huang:** Writing – review & editing, Methodology, Funding acquisition. **Guisong Liu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Sichuan Science and Technology Program, China (Grant No. 2023ZYD0145), the Natural Science Foundation of Sichuan Province, China (Grant No. 2023NSFSC1411 and No. 2025ZNSFSC1495).

Data availability

Data will be made available on request.

References

Aghli, N., & Ribeiro, E. (2021). Combining weight pruning and knowledge distillation for cnn compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3191–3198).

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*, 211–236. http://dx.doi.org/10.1257/jep.31.2.211.

Arcuri, M. C., Gandolfi, G., & Russo, I. (2023). Does fake news impact stock returns? Evidence from US and EU stock markets. *Journal of Economics and Business*, *125–126*, Article 106130. http://dx.doi.org/10.1016/j.jeconbus.2023.106130, URL https://www.sciencedirect.com/science/article/pii/S0148619523000231.

Arefeen, M. A., Debnath, B., & Chakradhar, S. (2024). LeanContext: Cost-efficient domain-specific question answering using LLMs. *Natural Language Processing Journal*, *7*, Article 100065. http://dx.doi.org/10.1016/j.nlp.2024.100065, URL https://www.sciencedirect.com/science/article/pii/S294971912400013X.

Bayes, T. (1968). Naive bayes classifier. *Article Sources and Contributors*, 1–9.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., et al. (2018). Mutual information neural estimation. In J. Dy, A. Krause (Eds.), *Proceedings of machine learning research*: *80*, *Proceedings of the 35th international conference on machine learning* (pp. 531–540). PMLR, URL https://proceedings.mlr.press/v80/belghazi18a.html.

Burt, D. R., Ober, S. W., Garriga-Alonso, A., & van der Wilk, M. (2020). Understanding variational inference in function-space. http://dx.doi.org/10.48550/arXiv.2011.09421, arXiv:2011.09421.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th international conference on world wide web* (pp. 675–684). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/1963405.1963500.

Chang, Q., Li, X., & Duan, Z. (2024). Graph global attention network with memory: A deep learning approach for fake news detection. *Neural Networks*, *172*, Article 106115. http://dx.doi.org/10.1016/j.neunet.2024.106115, URL https://www.sciencedirect.com/science/article/pii/S0893608024000297.

Chen, Z., Hui, S. C., Zhuang, F., Liao, L., Jia, M., Li, J., et al. (2024). A syntactic evidence network model for fact verification. *Neural Networks*, *178*, Article 106424. http://dx.doi.org/10.1016/j.neunet.2024.106424, URL https://www.sciencedirect.com/science/article/pii/S0893608024003484.

Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., et al. (2022). Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022* (pp. 2897–2905). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3485447.3511968.

Chen, C., & Shu, K. (2023). Can LLM-generated misinformation be detected?. http://dx.doi.org/10.48550/arXiv.2309.13788, ArXiv arXiv:2309.13788.

Chen, C., & Shu, K. (2024). Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine*, *45*(3), 354–368. http://dx.doi.org/10.1002/aaai.12188, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12188. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12188.

Chen, Y., Sui, J., Hu, L., & Gong, W. (2019). Attention-residual network with CNN for rumor detection. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1121–1130). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3357384.3357950.

Chen, X., Zhou, F., Trajcevski, G., & Bonsangue, M. (2022). Multi-view learning with distinguishable feature fusion for rumor detection. *Know.-Based Syst*, *240*(C), http://dx.doi.org/10.1016/j.knosys.2021.108085.

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–Decoder approaches. In *Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation* (pp. 103–111). http://dx.doi.org/10.3115/v1/W14-4012, URL https://aclanthology.org/W14-4012.

Cui, L., Seo, H., Tabar, M., Ma, F., Wang, S., & Lee, D. (2020). DETERRENT: Knowledge guided graph attention network for detecting healthcare misinformation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 492–502). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3394486.3403092.

Dai, B., Zhu, C., Guo, B., & Wipf, D. (2018). Compressing neural networks using the variational information bottleneck. In *International conference on machine learning* (pp. 1135–1144). PMLR.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

Dong, Y., He, D., Wang, X., Jin, Y., Ge, M., Yang, C., et al. (2024). Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. *38*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 8354–8362). http://dx.doi.org/10.1609/aaai.v38i8.28677, URL https://ojs.aaai.org/index.php/AAAI/article/view/28677.

Dun, Y., Tu, K., Chen, C., Hou, C., & Yuan, X. (2021). KAN: Knowledge-aware attention network for fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*, 81–89. http://dx.doi.org/10.1609/aaai.v35i1.16080.

Federici, M., Dutta, A., Forré, P., Kushman, N., & Akata, Z. (2020). Learning robust representations via multi-view information bottleneck. *CoRR*, http://dx.doi.org/10.48550/arXiv.2002.07017, arXiv:2002.07017.

Guan, J., Dodge, J., Wadden, D., Huang, M., & Peng, H. (2024). Language models hallucinate, but may excel at fact verification. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)* (pp. 1090–1111). Mexico City, Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.naacl-long.62, URL https://aclanthology.org/2024.naacl-long.62.

Gupta, S., Hoffman, J., & Malik, J. (2016). Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2827–2836).

Hassan, A., Qazvinian, V., & Radev, D. (2010). What's with the attitude? Identifying sentences with attitude in online discussions. In H. Li, & L. Màrquez (Eds.), *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1245–1255). Cambridge, MA: Association for Computational Linguistics, URL https://aclanthology.org/D10-1121.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28.

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. ArXiv arXiv:1503.02531 URL https://api.semanticscholar.org/CorpusID:7200347.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., et al. (2024). Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *38*, In *Proceedings of the AAAI conference on artificial intelligence* (pp. 22105–22113). http://dx.doi.org/10.1609/aaai.v38i20.30214.

Hu, Z., Wei, L., Li, K., Liu, Z., Wang, Y., & Zhang, X. (2024). LLM-driven external knowledge integration network for rumor detection. In *International conference on intelligent computing* (pp. 3–13). Springer, http://dx.doi.org/10.1007/978-981-97-5678-0_1.

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., et al. (2021). Compare to the knowledge: Graph neural fake news detection with external knowledge. In C. Zong, F. Xia, W. Li, R. Navigli (Eds.), *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)* (pp. 754–763). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2021.acl-long.62, URL https://aclanthology.org/2021.acl-long.62.

Huang, K.-H., McKeown, K., Nakov, P., Choi, Y., & Ji, H. (2023). Faking fake news for real fake news detection: Propaganda-loaded training data generation. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 14571–14589). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.815, URL https://aclanthology.org/2023.acl-long.815.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), http://dx.doi.org/10.1145/3703155.

Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia Tools and Applications*, *80*(8), 11765–11788. http://dx.doi.org/10.1007/s11042-020-10183-2.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, http://dx.doi.org/10.48550/arXiv.1412.6980, URL https://api.semanticscholar.org/CorpusID:6628106.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108). http://dx.doi.org/10.1109/ICDM.2013.61.

Li, G., Lu, W., Zhang, W., Lian, D., Lu, K., Mao, R., et al. (2024). Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. arXiv preprint arXiv:2403.09747.

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media Culture & Society*, *42*, http://dx.doi.org/10.1177/0163443720906992.

Ling, T., Chen, L., Lai, Y., & Liu, H.-L. (2024). Span-based few-shot event detection via aligning external knowledge. *Neural Networks*, *176*, Article 106327. http://dx.doi.org/10.1016/j.neunet.2024.106327, URL https://www.sciencedirect.com/science/article/pii/S089360802400251X.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: A robustly optimized BERT pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692, ArXiv arXiv:1907.11692. URL https://api.semanticscholar.org/CorpusID:198953378.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14–23. http://dx.doi.org/10.1038/nmeth.4370.

Lucas, J., Uchendu, A., Yamashita, M., Lee, J., Rohatgi, S., & Lee, D. (2023). Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 14279–14305). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.883, URL https://aclanthology.org/2023.emnlp-main.883.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K., et al. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence* (pp. 3818–3824).

Ma, J., Gao, W., & Wong, K.-F. (2019). Detect rumors on Twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference* (pp. 3049–3055). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3308558.3313741.

Ma, B., Lin, D., & Cao, D. (2017). Content representation for microblog rumor detection. In P. Angelov, A. Gegov, C. Jayne, & Q. Shen (Eds.), *Advances in computational intelligence systems* (pp. 245–251). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-46562-3_16.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*(3), 190–195. http://dx.doi.org/10.1177/0272989X8900900307.

Micaelli, P., & Storkey, A. J. (2019). Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, *32*.

Nan, Q., Cao, J., Zhu, Y., Wang, Y., & Li, J. (2021). MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 3343–3347).

Onan, A., & Alhumyani, H. A. (2024). DeepExtract: Semantic-driven extractive text summarization framework using LLMs and hierarchical positional encoding. *Journal of King Saud University - Computer and Information Sciences*, *36*(8), Article 102178. http://dx.doi.org/10.1016/j.jksuci.2024.102178, URL https://www.sciencedirect.com/science/article/pii/S1319157824002672.

Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., & Wang, W. (2023). On the risk of misinformation pollution with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the association for computational linguistics: EMNLP 2023* (pp. 1389–1403). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.findings-emnlp.97, URL https://aclanthology.org/2023.findings-emnlp.97.

Pan, L., Wu, X., Lu, X., Luu, A. T., Wang, W. Y., Kan, M.-Y., et al. (2023). Fact-checking complex claims with program-guided reasoning. In A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 6981–7004). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.386, URL https://aclanthology.org/2023.acl-long.386.

Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., et al. (2023). Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4. In H. Bouamor, J. Pino, K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6399–6429). Singapore: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.emnlp-main.395, URL https://aclanthology.org/2023.emnlp-main.395.

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). Declare: Debunking fake news and false claims using evidence-aware deep learning. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 22–32). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-1003, URL https://aclanthology.org/D18-1003.

Raj, C., & Meel, P. (2022). ARCNN framework for multimodal infodemic detection. *Neural Networks*, *146*, 36–68. http://dx.doi.org/10.1016/j.neunet.2021.11.006, URL https://www.sciencedirect.com/science/article/pii/S0893608021004342.

Rocha, Y. M., De Moura, G. A., Desidério, G. A., De Oliveira, C. H., Lourenço, F. D., & de Figueiredo Nicolete, L. D. (2021). The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. *Journal of Public Health*, 1–10. http://dx.doi.org/10.1007/s10389-021-01658-z.

Schlicht, I. B., Sezerer, E., Tekir, S., Han, O., & Boukhers, Z. (2021). Leveraging commonsense knowledge on classifying false news and determining checkworthiness of claims. http://dx.doi.org/10.48550/arXiv.2108.03731, ArXiv arXiv:2108.03731 URL https://api.semanticscholar.org/CorpusID:236956566.

Sheng, Q., Cao, J., Zhang, X., Li, R., Wang, D., & Zhu, Y. (2022). Zoom out and observe: News environment perception for fake news detection. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4543–4556). Dublin, Ireland: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2022.acl-long.311, URL https://aclanthology.org/2022.acl-long.311.

Sheng, Q., Zhang, X., Cao, J., & Zhong, L. (2021). Integrating pattern- and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM international conference on information & knowledge management* (pp. 1640–1650). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3459637.3482440.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 395–405). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3292500.3330935.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (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. http://dx.doi.org/10.1089/big.2020.0062.

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J. L., et al. (2022). Prompting GPT-3 to be reliable. ArXiv, arXiv:2210.09150.

Srivastava, A., Dutta, O., Gupta, J., Agarwal, S., & AP, P. (2021). A variational information bottleneck based method to compress sequential networks for human action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2745–2754).

Su, J., Cardie, C., & Nakov, P. (2024). Adapting fake news detection to the era of large language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Findings of the association for computational linguistics: NAACL 2024* (pp. 1473–1490). Mexico City, Mexico: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.findings-naacl.95, URL https://aclanthology.org/2024.findings-naacl.95.

Sun, M., Zhang, X., Zheng, J., & Ma, G. (2022). DDGCN: Dual dynamic graph convolutional networks for rumor detection on social media. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*, 4611–4619. http://dx.doi.org/10.1609/aaai.v36i4.20385.

Tian, X., Zhang, Z., Lin, S., Qu, Y., Xie, Y., & Ma, L. (2021). Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *2021 IEEE/CVF conference on computer vision and pattern recognition* (pp. 1522–1531). Los Alamitos, CA, USA: IEEE Computer Society, http://dx.doi.org/10.1109/CVPR46437.2021.00157, URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.00157.

Tishby, N., Pereira, F., & Bialek, W. (2001). The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, *49*, http://dx.doi.org/10.48550/arXiv.physics/0004057.

Trueman, T. E., J., A. K., P., N., & J., V. (2021). Attention-based C-bilstm for fake news detection. *Applied Soft Computing*, *110*, Article 107600. http://dx.doi.org/10.1016/j.asoc.2021.107600, URL https://www.sciencedirect.com/science/article/pii/S1568494621005214.

Vaibhav, V., Mandyam, R., & Hovy, E. (2019). Do sentence interactions matter? Leveraging sentence level representations for fake news classification. In D. Ustalov, S. Somasundaran, P. Jansen, G. Glavaˇ s, M. Riedl, M. Surdeanu, & M. Vazirgiannis (Eds.), *Proceedings of the thirteenth workshop on graph-based methods for natural language processing (textGraphs-13)* (pp. 134–139). Hong Kong: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D19-5316, URL https://aclanthology.org/D19-5316.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Red Hook, NY, USA: Curran Associates Inc..

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations*, Accepted as poster. URL https://openreview.net/forum?id=rJXMpikCZ.

Wang, J., Bao, W., Sun, L., Zhu, X., Cao, B., & Philip, S. Y. (2019). Private model compression via knowledge distillation. *33*, In *Proceedings of the AAAI conference on artificial intelligence* (01), (pp. 1190–1197).

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., et al. (2018). Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 849–857). http://dx.doi.org/10.1145/3219819.3219903.

Wu, L., Liu, P., Zhao, Y., Wang, P., & Zhang, Y. (2024). Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, *36*(1), 211–225. http://dx.doi.org/10.1109/TKDE.2023.3280555.

Wu, L., Rao, Y., Yang, X., Wang, W., & Nazir, A. (2020). Evidence-aware hierarchical interactive attention networks for explainable claim verification. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 1388–1394). International Joint Conferences on Artificial Intelligence Organization, http://dx.doi.org/10.24963/ijcai.2020/193, Main track.

Wu, K., Yang, S., & Zhu, K. Q. (2015). False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering* (pp. 651–662). http://dx.doi.org/10.1109/ICDE.2015.7113322.

Wu, S., Yu, J., Chen, J., & Zhou, W. (2024). Generative commonsense knowledge subgraph retrieval for open-domain dialogue response generation. *Neural Networks*, *180*, Article 106666. http://dx.doi.org/10.1016/j.neunet.2024.106666, URL https://www.sciencedirect.com/science/article/pii/S0893608024005902.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence, IJCAI-17* (pp. 3901–3907). http://dx.doi.org/10.24963/ijcai.2017/545.

Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., & Shu, K. (2021). Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021* (pp. 3465–3476). http://dx.doi.org/10.1145/3442381.3450004.

Zhang, X., & Gao, W. (2023). Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. A. Krisnadhi (Eds.), *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the Asia-Pacific chapter of the association for computational linguistics (volume 1: long papers)* (pp. 996–1011). Nusa Dua, Bali: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.ijcnlp-main.64, URL https://aclanthology.org/2023.ijcnlp-main.64.

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., et al. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. ArXiv arXiv:2309.01219 URL https://api.semanticscholar.org/CorpusID:261530162.

Zhang, C., & Peng, Y. (2018). Better and faster: knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. arXiv preprint arXiv:1804.10069.

Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., & Ma, K. (2019). Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3713–3722).

Zhang, Y., Xiang, T., Hospedales, T. M., & Lu, H. (2018). Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4320–4328).

Zhang, L., Zhang, X., Li, C., Zhou, Z., Liu, J., Huang, F., et al. (2024). Mitigating social hazards: Early detection of fake news via diffusion-guided propagation path generation. In *ACM multimedia*.

Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web* (pp. 1395–1405). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/2736277.2741637.

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), http://dx.doi.org/10.1145/3395046.

Zhu, Y., Sheng, Q., Cao, J., Li, S., Wang, D., & Zhuang, F. (2022). Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 2120–2125). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3477495.3531816.