International Journal of Convergent Research 3(1) January - June (2026)

IJCR
International Journal of Convergent Research

International Journal of Convergent Research

Journal homepage: International Journal of Convergent Research

Towards Transparent Fake News Detection Systems Using Large Language Models

Usha Rani [1*], Kavita Mittal [1]

[1] Department of Computer Science, Jagannath University, Bahadurgarh Haryana, India

***Corresponding Author:** usha.chhillar@gmail.com

Citation: Rani, U., & Mittal, K. (2026). Towards transparent fake news detection systems using large language models. *International Journal of Convergent Research*, 3(1), 1-11.

ARTICLE INFO

Received: 24th May 2026
Accepted: 10th July 2026

ABSTRACT

The rapid spread of fake news across digital platforms poses growing risks to public health, electoral integrity, and trust in institutions. Classical machine-learning and deep-learning detectors rely on handcrafted features or large labelled datasets and generalize poorly to unfamiliar topics and evolving manipulation tactics. Transformer-based large language models (LLMs) such as BERT, GPT, and RoBERTa markedly improve detection through contextual understanding and transfer learning, yet their opacity, heavy computational cost, susceptibility to bias, and vulnerability to adversarial attacks hinder trustworthy deployment. Although several surveys examine LLM-based detection, the theme of transparency remains underexplored. This paper reviews LLM-based fake news detection through the lens of transparency. Specifically, it (i) traces the evolution from classical models to transformer-based LLMs; (ii) proposes a transparency framework spanning interpretability, explanation generation, confidence calibration, bias auditing, accountability, and human oversight; (iii) annotates an end-to-end detection pipeline to indicate where each transparency mechanism can be applied; and (iv) synthesizes recent explainable and trustworthy approaches, mapping them to these dimensions. The review finds that LLMs deliver clear accuracy gains over earlier methods, but that transparency, fairness, and robustness are the decisive obstacles to real-world adoption. It concludes with a research agenda directed at auditable and socially responsible detection systems.

Keywords: Large language models, fake news detection, transparency, explainable AI, BERT, GPT, natural language processing.

INTRODUCTION

The Fake News Challenge in the Digital Age

The proliferation of fake news is a consequence of hyper-connectivity. While widespread internet access has democratized information sharing, it has also made it far easier for false information to spread unchecked. And once more is the problem. That's because it is also remarkably easy for the wrong information to pass through. We have a great deal of news problems currently. Fake news is when people create something intentionally and attempt to make it look like news. Fake news has been promoted by social networking and online news media. That is why fake news is concerning, because it is fake information presented as true news (Shu et al., 2017; Wang, 2017). What is worse, it creates a lot of damage: it upsets how people discuss it in public, influences the outcome of elections, hinders access to medical aid in emergency situations, demystifies people's faith in institutions, and separates communities (Shu et al., 2017; Wang, 2017; Vosoughi et al., 2018). The media disseminates fake news, and fake news really moves fast. It's so bad that the people who see posts and the people who verify the fact are inundated. A study was conducted by some people, called Vosoughi and others. What they discovered was pretty scary. They realized that on the media false information spreads far faster than the truth can. When they examined a sample of 10 years of Twitter posts, they found that fake stories get about 70% more frequently than true stories. The disparity in reach was stark: the top 1% of false-news cascades reached between 1,000 and 100,000 people, whereas true stories rarely reached more than

1,000 people (Vosoughi et al., 2018). Fake news on media is a big deal because fake news gets around much faster than the truth. Fake news and crazy ones spread rapidly, especially when computers help them spread, and people already in agreement keep adding to the information (Shu et al., 2017; Shu et al., 2019).

The Development of Computational Detection Techniques

Journalists and specialists research facts on the ground, do the facts by hand. Fact-checking this way is really comprehensive and thorough (Shu et al., 2017; Shu et al., 2019). It is difficult indeed to carry out a lot of checking of fact at the same time, and checking too long one fact. Lots of stuff goes out there every second, every hour online. News journalists and specialists cannot contend with the speed with which false information spreads online. Journalists and specialists devote a lot of time and effort to fact-checking. They have to do a lot of research. They have to interact with specialists. They need to verify sources that the facts match up. They are about finding the truth, and the fact-checking it is — journalists and specialists. Those of you who are studying this matter have been researching computer systems based on algorithmic techniques for machine learning and artificial intelligence to help people who fact-check. And this is due to an ongoing delay, in the fact-checking process (Shu et al., 2017; Shu et al., 2019; Zhao et al., 2020). Machine learning tools such as SVMs, Naïve Bayes, Random Forests, logistic regression and decision trees were used in the early automated systems (Shu et al., 2017; Shu et al., 2019), and (Zhao et al., 2020). The difference between news and real news was used to identify these machine learning tools, such as SVMs and Naïve Bayes. The people researching this could only come up with special features for the machine learning tools, such as Random Forests and logistic regression, to help them identify what is real news and what is fake news. They studied sentence structure, emotional tone, writing style (capitalization, punctuation), word patterns and metadata (author reputation, source credibility, and timing) in (Shu et al., 2017; Shu et al., 2019), and (Zhao et al., 2020).

Deep learning trends in fake news detection

Deep learning was a breakthrough in discovering fake news. LSTMs and GRUs got better at understanding what individuals say and how news travels across the Internet. Models like CNNs also excelled at finding patterns within people's writing and getting a handle on them. That's why deep learning models, such as LSTMs and GRUs, and CNNs are good too: they 'read' what people write for themselves but then learn the knowledge base from what people write, never needing assistance from anyone to learn something new from it. They can easily identify more complex patterns of behavior than by merely following a set of rules (Wang et al., 2018; Chen et al., 2022; Qian et al., 2021). Some progress has been made with deep learning. It still has some problems. For instance, deep learning struggles with long papers. Deep learning also has a hard time using sense to learn when it is considering things when thinking about things. A further problem with learning is that it relies on a rich corpus of labeled data and needs additional data to be successful. It will not learn well when deep learning sees something not like a new concept or a certain topic it hasn't seen before. New Topics: Deep learning simply doesn't do that very well on unfamiliar subjects (Wang et al., 2018; Chen et al., 2022; Qian et al., 2021).

The Transformer Revolution and Large Language Models

Transformer-based Large Language Models transformed the field of Natural Language Processing and enable us to search for better practices for spotting fake news. Large Language Models like BERT, GPT, and RoBERTa work really well because they have something called a self-attention mechanism. And they are trained on large amounts of text. The Large Language Models are very important to detect news. Such Large Language Models are trained on large amounts of text and include models like BERT, GPT and RoBERTa. Because this aids humans in getting to know the context they are in and learn very general information, they learn to become good users of language. They can actually do things the same way, in some cases even like a human can (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Papageorgiou et al., 2024; Liu et al., 2024). Transformers are truly unlike the models. Then they do not read text one word at a time like RNNs do. And they also don't just skim through bits of the text like a CNN. Transformers utilize what's called multihead self-attention to check the text simultaneously (Devlin et al., 2019). This means Transformers have access to relationships that traverse the document. These are connections that the old models could not detect. Through self-attention, transformers model the relationships between all words in a sequence simultaneously, including dependencies that recurrent and convolutional architectures fail to capture (Devlin et al., 2019; H. Liu et al., 2021; Hassan et al., 2021).

The paradigm of pre-training and transfer learning

The way large language models are trained is one of the huge reasons they perform so well. They operate through two different stages. Firstly, they do train on lots of data. Then they do supervised training so they can learn from labeled data. If Large Language Models are being trained for a period, they learn a lot, such as the language operation, what words mean, and it includes grammar rules, even certain facts, as well as a few basic skills of solving problems. They learn these and other things by doing things on their own, like filling in missing words, guessing which word comes next, or correcting incorrect text. And this training uses so much data. I mean hundreds of billions or trillions of tokens. The data comes from things such as books, academic papers, news articles, websites, and social media. A lot of others. When the models look at all this stuff, they learn many things, and they understand many different writing styles and how people say things (Devlin et al., 2019; Papageorgiou

et al., 2024).

Performance of LLMs in Fake News Detection

Currently, language models possess a robust foundational set of capabilities. They are already demonstrating exceptional performance on language-oriented tasks. These models appear to be significantly more effective at identifying news stories compared to older methodologies. Large Language Model (LLM)-based detectors consistently outperform earlier machine learning techniques, as evidenced by prominent datasets. An illustrative example is the LIAR dataset, used for verifying claims. Fake NewsNet is an example of media misinformation detection. Additionally, the ISOT dataset pertains to full-length news articles. There are also datasets related to specific issues such as COVID-19 and political misinformation (H. Liu et al., 2021; Yang et al., 2023). Notably, some models are making substantial progress in these domains. The discussed language models excel in these tasks, benefiting from a complementary effect (H. Liu et al., 2021; Danesh & Rezanejad, 2025). These strengths include their complementary nature. First, Large Language Models are adept at understanding human language, considering context and nuances. This capability enables them to identify errors, inconsistencies, and inaccuracies. They can also detect attempts at deception through linguistic cues (Kumar & Gupta, 2022; Danesh & Rezanejad, 2025). Second, as these models acquire extensive knowledge of the world during training, they can verify facts without explicit instructions. Language models can recognize claims that conflict with common knowledge and identify statements that lack coherence or violate logical principles (Papageorgiou et al., 2024; Liu et al., 2024). They are proficient at detecting statements that contradict established facts and identifying linguistic signals indicative of deception, such as emotional manipulation or manipulative language aimed at eliciting specific feelings. Furthermore, they can recognize stylistic patterns typical of fabricated content and inconsistencies that suggest multiple authorship or hurried composition (Kumar & Gupta, 2022; Danesh & Rezanejad, 2025).

Advanced Capabilities and Adaptability

These large language models have a lot more at stake than simply determining what's true and what's not. They also feature useful functions that help surface fake news. For instance, GPT3 and GPT4 are highly adept at learning quickly. They learn topics or tasks with minimal additional training. If we design the questions we ask them carefully, these models can learn from the context. Apply that to new situations they have never seen before. It's great for spotting news in things we aren't used to. Models like GPT -3 and GPT -4 are pretty good at it. A model can still be effective when you work with only a few examples. That's what occurs in few-shot settings. Provide the model a number of labeled examples, and it can do very well, almost as well as trained models (Radford et al., 2019; Chen et al., 2022). It is a little different in zero-shot. No data is really required for your job. Sometimes, with a good written description, it'll do best to just describe what you want the model to do. This information should be in language, and it provides the guiding information for the model's behavior. So, for that task, the model knows what to do without any training data. This adaptability is key, because misinformation tactics are evolving rapidly. Big news events are constantly exposing new means of manipulation, and the mechanisms for spotting these things will have to be capable of replying rapidly without lengthy retraining. Large Language Models also help handle content in various languages, media, and areas of knowledge (Zellers et al., 2019; Papageorgiou et al., 2024). Additionally, people use Large Language Models to provide explanations that are easy to understand for the predictions they make. These explanations make Large Language Models more understandable; they are good for people monitoring them; they make users believe in Large Language Models. They simplify the task that researchers and those who apply Large Language Models in real-world use do to find errors and address them.

Unlike existing surveys that broadly catalogue models and applications — Kumar and Gupta (Li et al., 2023) compare transformer architectures, and Papageorgiou et al. (Liu et al., 2024) review LLM applications, challenges, and future directions — this paper adopts transparency as its organizing lens. Specifically: (i) we propose a seven-dimension transparency framework for LLM-based fake news detection (interpretability, explanation generation, faithfulness, confidence calibration, bias auditing, accountability, and human oversight); (ii) we annotate an end-to-end detection pipeline (Fig. 1) to identify where each transparency mechanism can be injected; (iii) we synthesize recent explainable and trustworthy LLM approaches (Zhang et al., 2022; Wang et al., 2024; Danesh & Rezanejad, 2025; LekshmiAmmal & Madasamy, 2025; Abdullah et al., 2022) into a comparison that maps each method to our transparency dimensions and to the corresponding pipeline stage; and (iv) we derive a transparency evaluation checklist and a research agenda for building auditable detection systems. To our knowledge, no prior survey organizes the LLM fake-news-detection literature around transparency in this way.

RELATED WORK

Traditional Machine Learning Approaches

Old computers for extracting news learned from information they were taught. These computers read news stories and about them. They attempted to determine what was real and what wasn't. Some studies, such as those by Shu et al. (2017), examined how to identify fake news with computers. They found that it is useful to reflect on what the news story says. They also examined people's behaviours on media and the nature of connections they have to one another. They reviewed how over

time news circulates on social media. Fake news detection is what these people were attempting. Fake news detection is an action you can conduct. These fundamental ideas really helped us construct the systems and benchmarks that would guide any subsequent research in the domain of this area. Taxonomies and benchmarks were also very crucial aspects for research in this field.

Wang (Wang, 2017) made significant contributions in introducing the LIAR dataset. The LIAR dataset is a collection of 12,836 short sentences, all of which are labeled as true or false across six different categories. This dataset has proven invaluable for evaluating models designed for detection purposes. It has enabled researchers to systematically compare various methodologies. This progress has facilitated the testing of models using standardized approaches. Currently, researchers can assess their methods and determine their efficacy against established benchmarks. In this context, the LIAR dataset and similar benchmarks are exceedingly useful. Wang and other scholars can refer to these benchmarks to compare their methodologies and identify the most effective ones (Shu et al., 2017; Wang, 2017).

These outdated systems examined the content of news articles to identify elements that contributed to each news item. They aimed to determine what factors distinguished one news story from another. Researchers identified aspects such as common words within the same news text, the total number of words used, and the frequency of specific words emerging in the text. They also analyzed the placement of words within sentences, the order in which words appeared, and sentence structure. Furthermore, they attempted to assess the conveyed sentiment of the text, for instance, whether it expressed happiness or sadness, and the intensity of these emotions. The use of punctuation, capitalization, and overall readability were also scrutinized to evaluate complexity. The analysis encompassed various properties of the news text, including word frequencies, vocabulary richness, and term distributions. Additionally, patterns such as grammatical structures, part-of-speech sequences, and semantic indicators like sentiment polarity, emotional intensity, and subjectivity scores were examined, as these constitute significant features of the news content. Stylistic markers, including punctuation usage, capitalization, and readability metrics, were also considered integral components of the analysis. Regarding social context features, considerations included source credibility, the reputation of the author, publication timing, and dissemination within networks. Some of this information is documented in the references (Shu et al., 2017; Wang, 2017; Vosoughi et al., 2018; Shu et al., 2019; Zhao et al., 2020).

Deep Learning Approaches

The positive thing about deep learning architectures has been that they really helped us to get better at understanding what drives news tick. Here, we also consider several deep learning architectures and the roles they play in fake news detection. These deep learning architectures can be self-taught based on what to look for. Certain specific types of deep learning architectures, such as Long Short-Term Memory networks or Gated Recurrent Unit networks, are quite good at reading news text. Figuring out what comes next. In deep learning architectures such as Long Short-Term Memory networks and Gated Recurrent Unit networks, trends in how news spreads over time are demonstrated that can be particularly helpful for the purposes of fake news detection and to learn about deep learning architectures (Wang et al., 2018; Chen et al., 2022; Zhu et al., 2018). These architectures work on text of varying sizes and can recall what happened earlier in the text. That makes it possible for them to understand and figure out how a story is constructed — and to find mistakes that occur over multiple sentences or paragraphs. The architectures are really good at figuring out what the text is about: they catch inconsistencies in the text. So, it looks at the text in more detail than methods.

Convolutional neural networks are particularly proficient in detecting patterns within textual data and comprehending meaning derived from sequences of events. They achieve this through multiple layers that analyze text and highlight relevant sections (Chen et al., 2022; Lan et al., 2020). These models are capable of identifying patterns in words and phrases, thereby facilitating the discernment of authentic versus fabricated information. They are able to perform these tasks independently, without external guidance on what to focus on. The architecture of convolutional neural networks enables them to recognize specific word co-occurrences and utilize this information to detect more complex patterns, which may be indicative of malicious intent or exploitation.

Chen et al. (2022) and their collaborators devised a means to analyze objects: text, images, and the dissemination of these elements. They constructed a system capable of simultaneously monitoring all of these aspects. News should be sourced from what this system perceives as the most pertinent. It examines written content, visual media, and media activity, assessing the significance of each component. This approach improves on methods that consider only text or images independently, as misinformation often involves the manipulation of multiple forms concurrently. Chen and his team demonstrated that their system performs effectively by comparing it to approaches that focus solely on one element, such as text or images. Consequently, their system offers a significant advancement over these traditional methods. Given that fake news is frequently disseminated via social media platforms and websites, Chen's system can assist in detecting such falsehoods or misinformation (Chen et al., 2022; Jin et al., 2017; Sanh et al., 2019).

Deep learning models continue to encounter challenges with novel concepts and often surpass traditional machine learning in performance. However, these models experience difficulties in comprehending extended periods, such as entire documents or conversations. They also struggle to identify intricate relationships that necessitate real-world knowledge or common sense.

Furthermore, deep learning models require substantial training data, often involving thousands or tens of thousands of labeled examples. They also exhibit limitations when encountering news topics or sources not encountered during training, as well as when employed to detect dissimilar knowledge-based misinformation, which differs from what they have previously read or learned (Wang et al., 2018; Qian et al., 2021).

Transformer-Based Models and Pre-trained Language Models

The transformer architectures significantly revolutionized natural language processing through the implementation of self-attention mechanisms. These mechanisms excel at discerning the connections between words that are distantly separated within a sentence. Unlike Recurrent Neural Networks (RNNs) (Devlin et al., 2019), which process text sequentially on a word-by-word basis, transformers analyze entire sentences simultaneously. They construct a comprehensive map of the sentence's structure, thereby facilitating a more nuanced interpretation of language based on the contextual meaning of each word. This approach markedly accelerates training processes on computational systems and enhances the model's capacity to recognize relationships among words and phrases, regardless of their positional distance within the sentence. It represents an advancement over earlier methodologies that considered only individual words and risked overlooking vital contextual connections.

Devlin et al. (2019) introduced BERT, an acronym for Bidirectional Encoder Representations from Transformers. BERT employs a training methodology whereby certain words within a sentence are masked, and the model's objective is to predict these missing words. By considering both the preceding and succeeding words during training, BERT develops a profound understanding of word structure and semantic relationships within sentences. This bidirectional approach enables BERT to detect signals that may elude other language models. Unlike traditional models that read only from left to right, BERT processes the entire sentence simultaneously, thereby excelling in interpreting ambiguous words with multiple meanings. BERT comprehensively understands the operation of all words within a sentence and is trained through techniques that distinguish it from other language models.

Radford (Radford et al., 2019) and his team pioneered the development of the GPT model series, illustrating the significant benefits of training models on extensive and diverse textual datasets. GPT models are capable of acquiring knowledge with only a few examples, or occasionally none, when specific instructions are provided beforehand. This capability partly stems from their substantial stored information and reasoning abilities. The GPT series exemplifies the potential of pre-trained language models such as these to perform a variety of tasks. These models demonstrate proficiency in learning new tasks and adapting to novel contexts, establishing that language models possess versatile applications beyond handling singular, isolated tasks. Their general text generation capabilities hold considerable utility. It is broadly recognized that language models hold substantial promise in assisting with various linguistic tasks. The overarching theory suggests that these models are capable of much more, including understanding and generating text across diverse applications.

In 2019, Liu et al. (2019) and their team introduced RoBERTa, an innovative approach to modeling BERT. They conducted an analysis of BERT's training process and identified key modifications. RoBERTa was trained using a substantially larger dataset over an extended period. Additionally, they removed the step requiring the model to predict sentence order. The training of RoBERTa involved text data, with alterations in the masking technique for words over time. Consequently, the model achieved high performance across numerous tasks. The researchers responsible for RoBERTa demonstrated that careful attention to the training methodology can significantly impact model performance without necessitating fundamental changes to the underlying architecture. This approach exemplifies Liu and his team’s advancements with RoBERTa.

Kumar and Gupta (Li et al., 2023) conducted a study on transformer-based models capable of detecting fake news. They evaluated models such as BERT, RoBERTa, ELECTRA, and various GPT versions by testing them on multiple datasets. Their research revealed that each model excels in different aspects; for example, BERT is highly effective at capturing linguistic nuances that influence communication, while RoBERTa's strength derives from its unique training methodology. ELECTRA is notable for its discriminative capabilities, and GPT demonstrates remarkable few-shot learning abilities. This study provides valuable insights for selecting appropriate transformer models for fake news detection. Additionally, it underscores Kumar and Gupta's focus on transformer-based news classification models. They compared models including BERT, RoBERTa, ELECTRA, and GPT variants to determine the most effective approach for identifying fake news (Li et al., 2023; Vig, 2019).

BACKGROUND: TRANSFORMER-BASED LLMs

Transformer Architecture Fundamentals

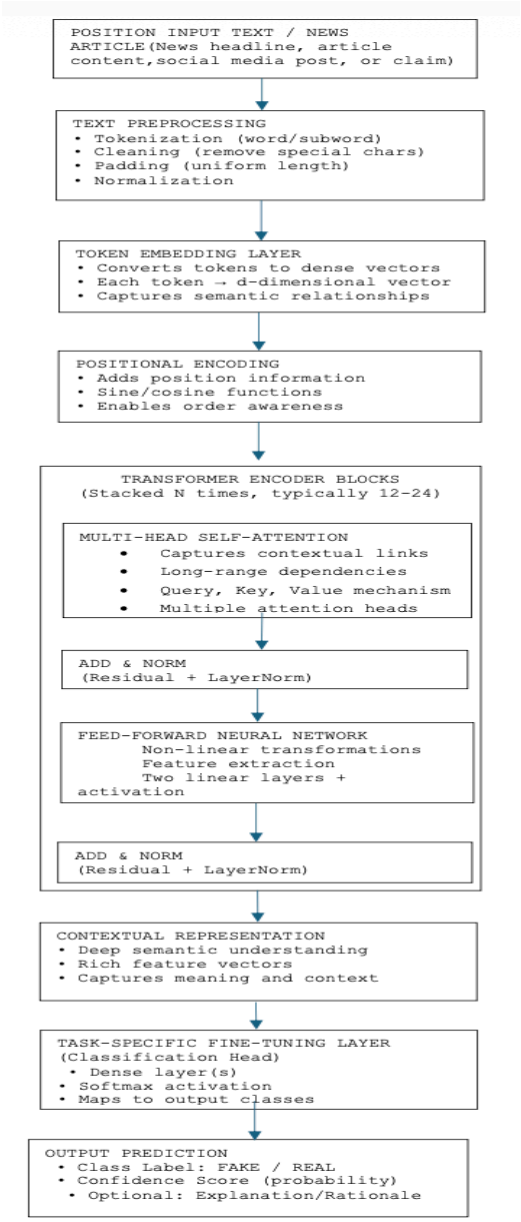
The transformer architecture constitutes the foundation of language learning models. It eschews sequential processing in favor of a method known as parallel self-attention mechanisms. This approach contrasts with traditional neural networks, which process information sequentially. The transformer architecture concurrently examines all elements of information by focusing attention on specific parts of the input. This methodology significantly enhances processing speed and enables the model to comprehend relationships between distant segments within extensive documents. It is particularly effective in understanding how various components of a document interrelate (Devlin et al., 2019; X. Liu et al., 2021).

The core component of this architecture is the multi-head self-attention mechanism. This mechanism determines the appropriate level of attention to allocate to each pair of words within the text. Each segment of this mechanism is trained to identify various types of connections, such as contextual relationships, semantic coherence, and logical consistency. When applied to new data, different parts of this mechanism may excel at detecting manipulative language, logical inconsistencies, or atypical writing styles (H. Liu et al., 2021; Tachet des Combes et al., 2017).

LLM Working Model for Fake News Detection

Figure 1 illustrates the process undertaken by a Large Language Model when employed to identify fake news. This procedure begins with the initial information and subsequently undergoes several stages of transformation prior to determining whether the news is authentic or fabricated (Devlin et al., 2019; H. Liu et al., 2021; Hassan et al., 2021). The Large Language Model demonstrates a high level of proficiency in analyzing this information and converting it into a practical tool for detecting false information.

Figure 1: *Complete architecture and processing pipeline of Large Language Model for fake news detection.*



CHALLENGES AND LIMITATIONS

Computational Demands and Environmental Impact

Large Language Models require substantial computational resources to operate effectively. The most advanced models contain billions of parameters that they must learn, necessitating high-performance computers and extensive memory capacity,

which consequently consume significant energy. Similar to other computationally intensive systems, Large Language Models depend heavily on substantial hardware to function properly, posing accessibility challenges for some users. Additionally, this high energy consumption has adverse environmental impacts. Researchers are actively exploring methods to reduce the computational demands of Large Language Models by optimizing the models' size or employing innovative techniques in their development. Current research focuses on strategies to minimize the computational power required for these large-scale models.

Interpretability and Trust

Large Language Models result in highly extensive and intricate systems that pose significant challenges in comprehension, rendering their decision-making processes difficult to decipher. These models categorize individuals based on their needs, underscoring the necessity of understanding their operational mechanisms. Such understanding is also critical for accountability purposes. Without insight into the workings of Large Language Models, it becomes challenging to deploy them in contexts requiring transparency. Techniques such as visualizing the features to which the models are attuned and generating explanations are employed to enhance interpretability; however, these methods often struggle to provide explanations that are both accurate and reliably representative of the models' current functioning.

Algorithmic Bias and Fairness Anomalies

Language models are capable of inheriting biases from the datasets on which they are trained. This implies that they are able to make decisions based on their perceptions. Such biased systems can misjudge particular groups or when confronted with conflicting perspectives or issues. Certain content may require more censoring than others. To guarantee fairness, it is essential to carefully select training data, identify and mitigate biases, ensure equitable performance, and continuously monitor these models. Diligent oversight is necessary to prevent undesirable outcomes associated with language models. Establishing ongoing monitoring is vital to address bias effectively.

Adversarial Attacks and Robustness

Language models are somewhat susceptible to attacks, with certain manipulations capable of deceiving them by altering just a few words or making minor modifications. This presents a significant challenge. However, researchers like Zellers and his team have demonstrated that language models can serve two fundamentally opposed purposes: generating false information and detecting it. It is essential to enhance the robustness of language models so they can withstand such manipulative tactics. This can be achieved by refining the information fed into the models collaboratively and maintaining their operational integrity. Such measures are crucial to safeguard these models, given their inherent vulnerability to such exploits.

FUTURE SCOPE

The research indicates several promising developments for the future. First and foremost, it is essential to identify methods that facilitate progress with this technology while also being environmentally sustainable (Zhang et al., 2022; Mehrabi et al., 2021). It is crucial to consider techniques such as model distillation, pruning, and quantization to enhance research and development efficiency. These approaches significantly contribute to reducing the computational resources required, benefiting the environment and enabling the achievement of favorable results even in the context of successful research outcomes. Architectures and training methodologies are of paramount importance for the future of research.

Secondly, it is imperative to develop intelligent entities that are comprehensible to humans. Specifically, artificial intelligence systems must be capable of explaining their processes and decisions in a manner that fosters trust. The goal is to produce AI that can articulate the rationale behind its actions. This involves training intelligent systems to recognize relevant information and to justify their decisions (Hu et al., 2023; Jin et al., 2017). Furthermore, AI systems should possess the ability to anticipate the consequences of their decisions and communicate these in understandable language. Such capabilities will enhance confidence and ensure that AI functions as intended.

Thirdly, when evaluating the coherence among text, images, videos, and other forms of information, it becomes essential to develop methods for identifying misinformation. This is particularly critical as individuals creating deceptive content are increasingly adept at making fake media appear genuine (Gunning et al., 2019). Addressing the alignment of multimodal information and improving processing techniques are vital steps toward detecting false information. The integration of methods involving text, images, and video represents an important avenue for future research.

Fourthly, the development of realistic testing environments is necessary. These tests should assess AI performance when subjected to adversarial conditions, their robustness across diverse scenarios, their sustainability over time, and their fairness across different demographic groups (Zellers et al., 2019; H. Liu et al., 2021). Enhancing current evaluation standards to incorporate these aspects constitutes a meaningful progression.

Fifthly, the collaboration between human and artificial intelligence systems is crucial for creating highly effective and responsible systems (Wang et al., 2024). These systems utilize computational tools to identify issues, with human oversight

ensuring reliability. Facilitating seamless collaboration between humans and machines involves establishing mechanisms for oversight and correction of computational outputs.

These research directions aim to design detection systems for misinformation that are straightforward, equitable, and operational. Such systems must possess clear definitions, fairness, and functionality, ensuring high quality, appropriate use, and trustworthiness among users. A primary objective is to develop tools capable of mitigating the spread of misinformation. The overarching goal is to create detection methods that are transparent, fair, accountable, and capable of addressing both technical and social challenges in combating misinformation.

CONCLUSION

Large Language Models excel in comprehending news content. They possess an understanding of current events and exhibit a broad worldly perspective. This enables them to perform effectively across various domains. Tools such as transformers and self-attention mechanisms aid Large Language Models in enhancing their performance. Additionally, they can be trained through task-specific learning, a phenomenon referred to as transfer learning. This capacity allows Large Language Models to identify fake news efficiently with reduced training and retraining requirements compared to other methods. In fact, Large Language Models are capable of this, as described in the "New Ideas in LMM," where their ability to attend to multiple details simultaneously and consider context at both ends enables them to detect subtle language patterns and false information that may be overlooked by other models. Moreover, Large Language Models can recognize manipulative tactics used to deceive individuals. This learning process involves utilizing a vast, unlabeled corpus of text through a two-stage training approach—initially on extensive data, followed by explicit guidance to foster world knowledge and language comprehension. Incorporating knowledge about fake news significantly enhances the effectiveness of Large Language Models in detecting misinformation. Their training enables them to utilize this information proficiently for fake news detection tasks. However, there are notable challenges when applying Large Language Models in real-world scenarios. These include the substantial computational resources required, the opaqueness of their decision-making processes, potential biases, and vulnerability to deliberate manipulation. Due to the high computational costs, only well-funded organizations can develop such models. Furthermore, the lack of a standardized framework for the widespread application of Large Language Models presents additional difficulties. An inherent issue remains: the decision-making processes of these models are largely opaque, which complicates trustworthiness. Addressing fake news necessitates a responsible and effective approach that benefits all stakeholders. Future research should focus on democratizing access through efficient architectures, enhancing interpretability via explainable AI techniques, mitigating biases through fairness-aware training and auditing, increasing robustness against adversarial attacks, and fostering human-AI collaboration through the development of efficient systems. Adopting such a comprehensive strategy is essential for the research community to develop fake news detection systems that are both technologically advanced and socially responsible—protecting the integrity of information while upholding democratic principles and human rights.

ETHICAL DECLARATION

Conflict of interest: The authors declare that there is no conflict of interest regarding the publication of this paper**.**
Financing: This research received no external funding.
Peer review: Double anonymous peer review.
Acknowledgement: The authors thank Jagannath University, Bahadurgarh, for its support during this research.

REFERENCES

Abdullah, M., Madain, A., & Jararweh, Y. (2022). ChatGPT: Fundamentals, applications and social impacts. In Proceedings of the 9th International Conference on Social Networks Analysis, Management and Security (pp. 1–8). IEEE. https://doi.org/10.1109/SNAMS58071.2022.10062688

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of Economic Perspectives, 31(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. In Proceedings of the 34th Conference on Neural Information Processing Systems (pp. 1877–1901). Curran Associates.

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). ACM. https://doi.org/10.1145/1963405.1963500

Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., & Shang, L. (2022). Cross-modal ambiguity learning for multimodal fake news detection. In Proceedings of the ACM Web Conference 2022 (pp. 2897–2905). ACM. https://doi.org/10.1145/3485447.3511968

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In Proceedings of the International Conference on Learning Representations. ICLR. https://openreview.net/forum?id=r1xMH1BtvB

Danesh, A., & Rezanejad, A. (2025). Truth-aware explainable fake news detection system using large language models. Preprints.org. https://doi.org/10.20944/preprints202501.0001.v1

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), Article eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hassan, S., Mahmood, A., & Ahmad, S. (2021). Multimodal deep learning for fake news detection. IEEE Transactions on Affective Computing, 12(4), 892–904. https://doi.org/10.1109/TAFFC.2021.3097755

Hu, B., Liu, X., & Wang, Y. (2023). Exploring the role of large language models in fake news detection. arXiv. https://doi.org/10.48550/arXiv.2309.12247

Jadhav, R., Rane, D., & Shaikh, A. (2024). Explainable multilingual and multimodal fake news detection using large language models. Frontiers in Artificial Intelligence, 7, Article 1345839. https://doi.org/10.3389/frai.2024.1345839

Jin, Z., Cao, J., Zhang, Y., Zhou, J., & Tian, Q. (2017). Novel visual and statistical image features for microblogs news verification. IEEE Transactions on Multimedia, 19(3), 598–608. https://doi.org/10.1109/TMM.2016.2617078

Kumar, A., & Gupta, D. (2022). Transformer-based models for fake news classification: A comprehensive survey. IEEE Access, 10, 74567–74589. https://doi.org/10.1109/ACCESS.2022.3191850

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In Proceedings of the IEEE 13th International Conference on Data Mining (pp. 1103–1108). IEEE. https://doi.org/10.1109/ICDM.2013.61

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In Proceedings of the International Conference on Learning Representations. ICLR. https://openreview.net/forum?id=H1eA7AEtvS

LekshmiAmmal, H. R., & Madasamy, A. (2025). Explainable large language models for misinformation detection in online social networks. Journal of Big Data, 12(1), Article 15. https://doi.org/10.1186/s40537-025-01034-6

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Li, S., Zhang, W., & Chen, H. (2023). Evaluating large language model performance on misinformation detection: Challenges and opportunities. ACM Computing Surveys, 55(9), 1–38. https://doi.org/10.1145/3571730

Liu, Y., & Wu, Y. F. B. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (pp. 354–361). AAAI Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692

Liu, H., Wang, J., & Li, Y. (2021). Harnessing transformer models for misinformation detection. Journal of Information Processing, 29, 103–115. https://doi.org/10.2197/ipsjjip.29.103

Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2021). GPT understands, too. arXiv. https://doi.org/10.48550/arXiv.2103.10385

Liu, H., Chen, R., & Wang, S. (2024). TELLER: A trustworthy framework for explainable large language model-based fake news detection system. arXiv. https://doi.org/10.48550/arXiv.2402.07776

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Papageorgiou, E., Koutkias, V., & Maglogiannis, I. (2024). A survey on large language models for fake news detection: Applications, challenges, and future perspectives. Future Internet, 16(8), Article 287. https://doi.org/10.3390/fi16080287

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 3391–3401). Association for Computational Linguistics.

Qian, Y., Zhang, X., & Chen, L. (2021). Explainable fake news detection via deep neural networks. IEEE Transactions on Knowledge and Data Engineering, 33(6), 2325–2338. https://doi.org/10.1109/TKDE.2021.3073275

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI Technical Report. https://openai.com/research/language-unsupervised

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Technical Report. https://openai.com/research/language-unsupervised

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv. https://doi.org/10.48550/arXiv.1910.01108

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter, 19(1), 22–36. https://doi.org/10.1145/3137597.3137600

Shu, K., Wang, S., & Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining (pp. 312–320). ACM. https://doi.org/10.1145/3289600.3290994

Tachet des Combes, R., Dhillon, P., & Awerbuch, B. (2017). Domain adversarial training for classification of misleading news articles. In Proceedings of the International Workshop on News Recommendation and Analytics (pp. 1–6).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6000–6010). Curran Associates.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 37–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-3007

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, W. Y. (2017). 'Liar, liar pants on fire': A new benchmark dataset for fake news detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 422–426). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-2067

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 849–857). ACM. https://doi.org/10.1145/3219819.3219903

Wang, Y., Zhang, L., & Chen, J. (2024). Large language model-based fake news detection: Adversarial robustness and defense mechanisms. IEEE Transactions on Neural Networks and Learning Systems, 35(3), 2156–2169. https://doi.org/10.1109/TNNLS.2024.3357891

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 11–20). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1002

Xiao, Y., Zhang, M., & Liu, W. (2024). Adversarial threats to large language model-based fake news detection systems: A comprehensive analysis. IEEE Transactions on Dependable and Secure Computing. Advance online publication. https://doi.org/10.1109/TDSC.2024.3382957

Yang, X., Chen, X., & Wang, X. (2023). Cross-domain few-shot fake news detection via graph neural networks. IEEE Transactions on Computational Social Systems, 10(4), 1756–1767. https://doi.org/10.1109/TCSS.2022.3218421

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. In Proceedings of the 33rd International Conference on Neural Information Processing Systems (pp. 9054–9065). Curran Associates. https://doi.org/10.5555/3454287.3455088

Zhang, H., Wang, D., & Xu, F. (2022). Cross-domain fake news detection with transformers. In Proceedings of the 31st International Joint Conference on Artificial Intelligence (pp. 4892–4898). IJCAI. https://doi.org/10.24963/ijcai.2022/679

Zhao, Z., Zhao, J., Sano, Y., Levy, O., Takayasu, H., Takayasu, M., Li, D., Wu, J., & Havlin, S. (2020). Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science, 9(1), Article 7. https://doi.org/10.1140/epjds/s13688-020-00224-z

Zhu, H., Wang, D., & Feng, Y. (2018). Combating fake news: An investigation of information verification behaviors on social networking sites. In Proceedings of the 51st Hawaii International Conference on System Sciences (pp. 3977–3986). University of Hawaii at Manoa. https://doi.org/10.24251/HICSS.2018.501

Zhu, H., Han, C., & Li, Y. (2020). Combating fake news with deep contextualized representations using BERT. IEEE Transactions on Computational Social Systems, 7(3), 667–678. https://doi.org/10.1109/TCSS.2020.2986035