

Zero Trust Frameworks for Misinformation Detection in Industry 5.0

Hirak Mazumdar ^{1,*}, Kamil Reza Khondakar ², Koushik Mukhopadhyay ¹, Sitikantha Chattapadhyay ³, Debdutta Pal ¹, Subhra Prokash Dutta ¹, Cheng-Chi Lee ^{4,5,*}

- 1 Department of Computer Science and Engineering, Adamas University, School of Engineering and Technology & Center of Excellence in Artificial Intelligence (COE-AI), Kolkata, 700126, India.
 - 2 School of Technology, Woxsen University, Hyderabad, 502345, India
 - 3 Department of Computer Science and Engineering, Techno International New Town, Kolkata, 700156, India
 - 4 Department of Library and Information Science, Artificial Intelligence Development Center, Fu Jen Catholic University, New Taipei City, 24205, Taiwan
 - 5 Department of Computer Science and Information Engineering, Fintech and Blockchain Research Center, Asia University, Taichung City, 41354, Taiwan
- * Correspondence author: hirak.mazumdar1@adamasuniversity.ac.in, cclee@blue.lins.fju.edu.tw

Abstract: This paper examines the application of the Generative Pre-trained Transformer (GPT) model for misinformation detection as part of Zero-Trust security models in Industry 5.0. As the transition to Industry 5.0 gains momentum, characterized by heightened interconnectedness and human-machine interaction, the necessity for rigorous security measures becomes critical. Disinformation poses significant threats, potentially undermining operational efficiency and safety, while causing substantial financial and reputational damage. To address these risks, we propose leveraging GPT, a natural language processing tool, to develop an effective misinformation detection system that enhances data integrity and trust in human-machine interactions within Industry 5.0. This paper discusses integrating a GPT-based misinformation detection system with existing security measures in Industry 5.0, showcasing its potential to strengthen a Zero-Trust security framework. We provide a detailed architectural specification of how the GPT-3 classifier integrates as a Policy Decision Point (PDP) within a Zero Trust Network Access (ZTNA) framework. Through case studies and simulations, we demonstrate the value of this machine learning model in protecting Industry 5.0 environments from misinformation-related threats. We also present inference latency analysis across GPT-3 variants to evaluate suitability for real-time industrial deployment. The research emphasizes the need for continuous vigilance and adaptable security strategies, highlighting the crucial role of GPT-based misinformation detection in maintaining a secure and reliable environment for human-machine interactions within the rapidly evolving context of Industry 5.0.

Keywords: Industry 5.0, Zero Trust, Misinformation Detection, GPT, Machine Learning, Data Integrity

1. Introduction

Generative Artificial Intelligence (GenAI) refers to a class of artificial intelligence (AI) systems that have the capability to generate new and original content. These systems use various techniques, often based on machine learning, to produce data or content that is not explicitly programmed but is rather learned from patterns and examples in the data it has been trained on. GenAI algorithms are widely used in consumer electronics such as voice assistants, smart speakers, wearables, and chatbots. Integrating with GenAI improves the systems as they can be more human-like responses, can adapt to user behavior and responses.

Although GenAI has several advantages they have a significant drawback of generating misinformation. Electronics like Alexa, voice-controlled assistants, chatbots, etc. can generate information that is incorrect. This can decrease the quality of consumer experience with electronic or software products.

Industry 4.0 aims towards technology innovations whereas Industry 5.0 aims towards value-driven

initiatives[1]. Moreover, Industry 5.0 aims towards human centric, sustainable and resilient[2,3] whereas Industry 4.0 might not be human-centric[1]. Specifically, Industry 4.0 implementations often prioritized full automation over human welfare; for instance, fully automated assembly lines frequently displaced workers without providing adequate upskilling or reskilling pathways. Data-driven decision systems in Industry 4.0 operated as opaque black boxes with no human-in-the-loop oversight, leaving workers unable to understand, challenge, or correct automated decisions affecting production quality and safety. Furthermore, worker well-being and ergonomics were frequently treated as secondary concerns to production optimization metrics in many Industry 4.0 deployments. A zero trust security approach can make the GenAI more safer to use and human centric. As zero trust security involves monitoring of each response automatically. This helps in making the systems better in terms of reducing misinformation.

To enable zero trust security in GenAI systems, it requires a misinformation detection algorithm to continuously monitor the responses and helps in generating correct information. The GPT-3 is found to be superior performance in understanding complex patterns. Additionally, it is one of the largest language model with 175 billions of parameters. Moreover, the misinformation can be highly dynamic and changes with time. The GPT-3 models can be reused via transfer learning making the models sustainable, which is a requirement of industry 5.0. This eventually reduces the time and costs.

Industry 5.0 emphasizes a shift towards human-centric manufacturing, where collaboration between humans and machines is optimized for sustainability and resilience. Our framework directly supports these goals by enhancing the security and trustworthiness of human-machine interactions. Through a meticulous design that incorporates Zero-Trust principles and advanced AI, our system ensures that human operators can interact with machines in a secure and efficient manner, thereby advancing the core objectives of Industry 5.0.

Finally, choosing GPT-3 models for the purpose of zero trust security by the detection of disinformation can improves the consumer experience in human machine interaction by reducing the generation of false information in consumer electronics or products in which GenAI is used. Moreover, it is human centric and sustainable which contributes towards industry 5.0.

Certainly, the key contributions of this paper are as follows:

1. As of our knowledge, we are the first to propose zero trust security through GPT-3 based disinformation detection for GenAI systems towards industry 5.0.
2. Highlighting the importance of misinformation detection in industry 5.0.
3. Proposing GPT-3 finetuning as zero trust security system for GenAI based systems for human centric and sustainable. That can be easily deployed in consumer electronics like voice assistants, chatbots, etc for better human machine interactions.
4. Through a series of simulations and case studies, this paper provides concrete examples of how a GPT-based disinformation detection system can contribute to the security of Industry 5.0 environments.
5. A detailed architectural integration of GPT-3 as a Policy Decision Point within a Zero Trust Network Access framework, bridging the gap between NLP classification and formal security architectures.
6. Inference latency analysis across GPT-3 variants (Ada, Babbage, Curie, Davinci) to evaluate real-time deployment feasibility in Industry 5.0 edge environments.

The rest of the manuscript is as follows: Section II describes recent literature and innovations in misinformation and Industry 5.0. Section III explains the proposed methodology. Section IV discusses the experimentation results, section V presents practical application and limitations and section VI provides conclusion.

2. Literature Survey

Misinformation has emerged as a significant challenge in the modern era, where advanced technologies and interconnected systems create new avenues for the dissemination of false information. This literature review explores key works in the domains of misinformation detection, language processing, and the evolving landscape of Industry 5.0.

He et al.[4] employ reinforcement learning in a case study focused on COVID-19 vaccine misinformation. Their work delves into the generation of counter-misinformation responses, demonstrating the applicability of reinforcement learning techniques in addressing real-world misinformation challenges. Dong et al.[5]

enhance mixup-based graph learning for language processing, providing insights into advanced techniques applicable to language-related challenges. While not explicitly addressing misinformation, their work on graph learning lays the foundation for understanding the intricacies of language representation.

Cui et al.[6] propose SAME, a sentiment-aware multi-modal embedding for detecting fake news. Though not directly related to Industry 5.0, their approach contributes valuable insights into the role of sentiment analysis in identifying deceptive information. Furukawa et al.[7] explore fake news detection through the analysis of biased user profiles in social networking sites. Their investigation sheds light on the human-centric aspects of misinformation, providing a socio-technical perspective on the challenges posed by deceptive information. Salama et al.[8] present CoAID-DEEP, an optimized intelligent framework for automated detection of COVID-19 misinformation on Twitter. This work emphasizes the role of intelligent systems in addressing the evolving landscape of misinformation in the context of global health crises.

Khan et al.[9] propose a Pyramidal Co-Attention Network for explainable detection of fake news. Their work aligns with the growing interest in interpretable models for misinformation detection. Canhasi et al.[10] address the lack of fake news corpora for under-resourced languages, focusing on Albanian fake news detection. Their work highlights the importance of considering linguistic diversity in misinformation detection. Jain et al.[11] present a comparative analysis of fake news classifiers on a low-resource Bengali language dataset. Their work contributes to understanding the challenges and opportunities in detecting misinformation in languages with limited resources. Shahid et al.[12] discuss challenges and future research opportunities in detecting and mitigating the dissemination of fake news.

Mourtzis et al.[13] explore blockchain integration in the industrial metaverse, emphasizing its role in ensuring data validity protection and organizational communication. Huang et al.[3] present a comparative analysis of Industry 5.0 and Society 5.0, addressing the sustainable and human-centric nature of these paradigms. Xu et al.[1] discuss the transition from Industry 4.0 to Industry 5.0, emphasizing the shift from technology-driven to value-driven approaches. Leng et al.[2] provide a comprehensive review of Industry 5.0, emphasizing its human-centric, sustainable, and resilient characteristics.

In recent years, the application of GPT models to misinformation detection has garnered significant attention. However, this study seeks to bridge a gap in the existing literature by uniquely integrating GPT into Zero-Trust security frameworks within the Industry 5.0 context. Unlike traditional applications, our approach not only utilizes GPT for real-time misinformation detection but also embeds it within a human-centric, sustainable, and resilient framework, aligning with Industry 5.0 principles. This research contributes new insights into how advanced AI models can be effectively leveraged to enhance security and trust in increasingly complex human-machine interactions. Moreover, contributing towards industry 5.0 by enabling zero trust security through misinformation detection for GenAI systems is unexplored as shown in Table 1.

Table 1. Comparison with State of the Art literature

Authors	Contribution	Methodology
Cui et al. [6]	Proposed a framework to integrate various features for fake news and significant improvement in the performance	Multi-modal embedding with sentiment awareness in fake news detection
Khan et al. [9]	Proposed a novel approach for explainable fake news detection	Using pyramidal co attention networks
Shadi et al. [14]	Improving the performance of the fake news detection	Enhancing BERT through attention mechanisms
Chen et al. [15]	Exploring detection of fake news on Covid-19	Explored various deep learning algorithms
Wazid et al. [16]	Proposed an intrusion detection machine learning model for healthcare applications in industry 5.0	Involves ensembling approaches through various machine learning classifiers
Li et al. [17]	Proposed STAR-RIS-CR-NOMA in Consumer IoT Networks for Industry 5.0 Resilience	Presented a novel approach named STAR-RIS-NOMA

Hussein et al. [18]	Introducing a novel CSI compression model that approximates sufficient statistics functions for individual channel matrices for industry 5.0	High degree of synchronization through the base station (BS) to obtain instant channel state information (CSI)
He et al. [4]	Proposed a reinforcement learning approach for the detection of the fake news in social media	A reinforcement framework via counter responses generation
Ours	AI based zero trust security for GenAI systems in industry 5.0 that is human centric and sustainable	Employing GPT-3 for misinformation detection

3. Methodology

This section outlines the methodology used for implementing the Generative Pre-trained Transformer (GPT-3) model for fake news detection, baselines and datasets used.

3.1. Data Collection

The models are trained using two datasets of fabricated news: GossipCop and PolitiFact. These datasets are sourced from FakeNewsNet [19–21].

In this manuscript, we introduce The GossipCop dataset consists of news articles centered around celebrity gossip. It comprises two files: *gossipcopreal.csv* and *gossipcopfake.csv*. The first file, *gossipcopreal.csv*, contains over 11,357 entries with accurate information, while the *gossipcopfake.csv* file includes over 10,783 entries with false information. These two files are merged and labeled as 1 for real news and 0 for fake news. The dataset includes columns such as id, URL, title, tweet-id, and true. For our purposes, we utilize the title column as the textual data[19–21].

Also, The PolitiFact dataset consists of news articles and tweets related to political news in the United States. It consists of two files: *politifactreal.csv*, which contains 432 tweets with genuine news, and *politifactfake.csv*, which includes 618 tweets with fabricated news. Similar to the GossipCop dataset, these files are merged and labeled as 1 for real news and 0 for fake news. The dataset contains columns such as id, URL, title, tweet id, and true[19–21].

3.2. Dataset Preparation

80% of the data is allocated for training purposes, while the remaining 20% is reserved for testing. The training data is utilized to train the machine learning model by providing input and output information. Once the model is trained, the training data becomes unseen data and is utilized to evaluate the model's performance.

The data preprocessing in this study encompasses text lowercasing, tokenization, removal of punctuation, and elimination of stopwords. Additionally, stemming and lemmatization are applied to reduce words to their base forms. In GPT finetuning, labels are transformed into text format to match the text-to-text translation nature of the task. This alignment ensures smooth handling of textual input and output, thereby improving the model's ability to produce accurate and contextually appropriate translations.

It is important to note that the above preprocessing steps (lowercasing, stopword removal, stemming, and lemmatization) were applied exclusively to the traditional machine learning baselines (GB, XGBoost, LR, DT, SVC), which require explicit feature engineering through TF-IDF vectorization. For all transformer-based models (BERT, DistilBERT, DeBERTa) and GPT-3 variants, the raw text was used with each model's native tokenizer (WordPiece for BERT variants; BPE for GPT-3). No stopword removal, stemming, or lemmatization was applied to transformer inputs, as these operations would disrupt the subword tokenization patterns learned during pre-training.

During the detailed finetuning process of the GPT-3 model, an important step is optimizing the input for better performance. This is done by appending a specific suffix, `\n\nIntent:\n\n`, to the input text. This additional sequence acts as a clear marker, indicating the end of the input section and ensuring the model processes the data correctly. Similarly, for the output generated by the finetuned GPT-3 model, it is crucial

to indicate the endpoint clearly. This is achieved by appending the output with the terminating tag END. This marker guides the model during the generation process, helping it produce output that meets the specified requirements and maintains coherence.

By using these optimization techniques during finetuning, the GPT-3 model can effectively utilize the `\n\nIntent:\n\n` suffix for input delimitation and the END tag for precise output termination. These careful measures enhance the model's fluency and accuracy, enabling it to generate contextually relevant and coherent outputs. Without these suffixes, the input and output tokens would not be distinguishable, leading to merged tokens, noise, and repetitive sentences in the model's output.

3.3. Baseline Models

In the experimental evaluation for fake news detection, several models have been utilized as baselines, each representing a distinct category of machine learning and deep learning approaches. These baseline models provide a comparative benchmark against which the performance of the GPT model variants can be assessed.

The baselines are opted based on the recent improvements and well-known algorithms for classification and NLP tasks. The machine learning models are foundational models, deep learning models and transformer models are considered to be state-of-the-art models in the NLP domain. Additionally, they are employed in the recent literature[14,15,22]. The first category includes traditional machine learning algorithms such as Gradient Boosting (GB)[23], XGBoost[24], Logistic Regression (LR)[25], Decision Trees (DT)[26], and Support Vector Classifier (SVC)[27], known for their interpretability, ease of implementation, and strong performance in various tasks. The second category comprises deep learning models, specifically Long Short-Term Memory (LSTM)[28], Bidirectional LSTM (Bi-LSTM)[29], and Convolutional Neural Network LSTM (CNN Bi-LSTM)[30], which excel at capturing complex patterns in sequential data, making them particularly effective for text-related tasks like fake news detection. The third category encompasses transformer-based models, including Bidirectional Encoder Representations from Transformers (BERT)[31], DistilBERT[32], and DeBERTa[33], which utilize self-attention mechanisms to understand word context and handle long-term dependencies, achieving state-of-the-art performance in numerous natural language processing tasks.

3.4. GPT-3 Finetuning

The Generative Pretrained Transformer 3 (GPT-3)[34] model is characterized by four distinct variants, namely Ada, Babbage, Curie, and Davinci. Ada, with the fewest parameters, represents the lower end of the complexity spectrum, while Davinci, embodying the most parameters, signifies the upper bound. Fine-tuning GPT-3, inherently a text-to-text transformer, for binary classification mandates a transformation of numeric labels into text-based equivalents. Traditional baseline models employ a numerical system with 0 denoting fake news and 1 signifying genuine news. For the fine-tuning process with GPT-3, these labels are transcribed into textual form: genuine news is labelled as "This is true news" and fake news is labelled as "This is fake news".

Upon the completion of labelling, a specific suffix, denoted by `\n\nIntent:\n\n`, is affixed to the input, and another suffix, END, is affixed to signify the output's end. The `\n\nIntent:\n\n` suffix was chosen following OpenAI's fine-tuning best practices documentation, which recommends a consistent, distinctive separator that does not commonly appear in natural text. The double newline characters ensure clear visual and tokenization-level separation between the input prompt and the expected classification output. Alternative delimiters such as `\n\nClassification:\n\n`, `###`, and `--->` were considered during preliminary experiments; however, the `\n\nIntent:\n\n` format was selected for consistency with OpenAI's recommended fine-tuning templates and because it yielded the most stable convergence during early training trials. These suffixes are of paramount importance, serving as clear delimiters for the input and output boundaries as shown in Figure 1.

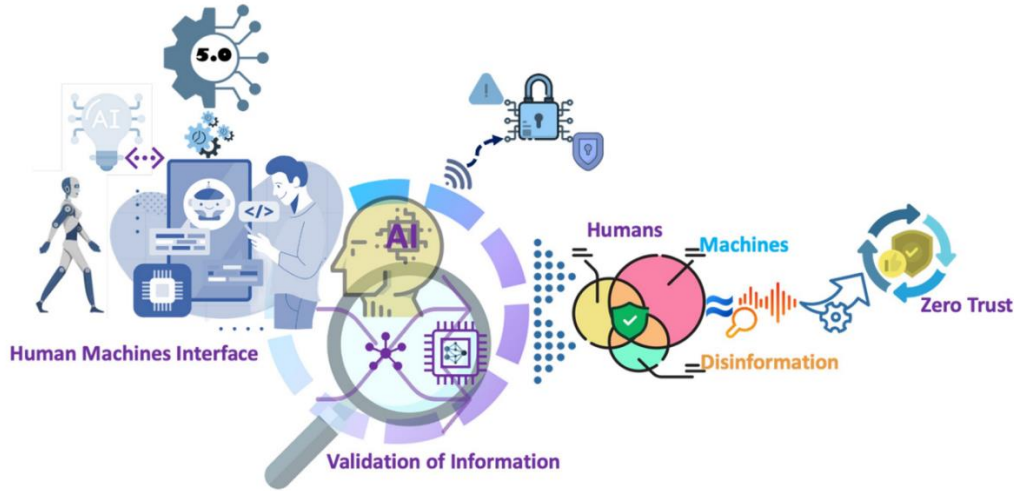


Figure 1. GPT-3 Finetuning

Thereafter, this data undergoes a transformation into a JSON format with prompt acting as the input and completion embodying the output. The culmination of this process involves fine-tuning the model on OpenAI servers, a process initiated by the execution of the command `openai api fine_tunes.create -t <TRAIN_FILE_ID_OR_PATH> -m <BASE_MODEL>`. The hyperparameters of the GPT-3 models include number of epochs, the maximum number of tokens, learning rate, batch size and frequency penalty. The number of epochs is 4, maximum combined prompt-plus-completion token length 2,049 and all other parameters are set to default.

Given that the input texts, in our datasets are typically short (10–30 tokens) and the classification completions are brief (4–5 tokens), the effective token usage per sample is well within the 2,049-token combined limit. The `max_tokens` parameter for completion generation was set to 10 to restrict output length and prevent extraneous generation.

The OpenAI Completions API returns log-probabilities (logprobs) for each generated token during inference. When the fine-tuned GPT-3 model generates a classification completion, the confidence score is computed by exponentiating the mean log-probability across all completion tokens: $\text{Confidence} = \exp(1/N \sum_i^N \log prob_i)$, where N is the number of tokens in the completion. This yields a continuous value between 0 and 1. Additionally, by setting the `logprobs` parameter to return probabilities for alternative tokens, the system compares the likelihood assigned to the first divergent token between the “true” and “fake” completions to derive a normalized binary confidence score. This confidence score is used by the PEP decision thresholds described in Section 2.5.

3.5. Integration with Zero Trust Architecture

To bridge the gap between the GPT-3 classification model and a formal Zero Trust security framework, we describe the architectural integration of the misinformation detection module within a Zero Trust Network Access (ZTNA) framework. In our proposed architecture, the GPT-3 fine-tuned classifier operates as a Policy Decision Point (PDP). Every response generated by a GenAI system (e.g., a chatbot, voice assistant, or automated report generator) is treated as an untrusted entity following the “never trust, always verify” principle.

The data flow proceeds as follows: (1) A GenAI system generates a response. (2) The response is intercepted by the Policy Enforcement Point (PEP) before delivery to the end user. (3) The PEP forwards the response text to the GPT-3 classifier (PDP) for misinformation assessment. (4) The GPT-3 classifier returns a classification label (real or fake) along with a confidence score. (5) Based on the classification output, the PEP enforces one of three actions: Allow (classification: real, confidence > 0.9), Flag for Human Review (confidence between 0.7 and 0.9), or Block (classification: fake, confidence > 0.9). (6) All decisions are logged in a centralized audit system for continuous monitoring and compliance.

This continuous verification loop ensures that no GenAI output is implicitly trusted, aligning with Zero Trust principles. The PEP can be deployed at the API gateway level for cloud-based GenAI services, or as an on-device middleware component for edge deployments in Industry 5.0 environments. Figure 2 illustrates this integration architecture.

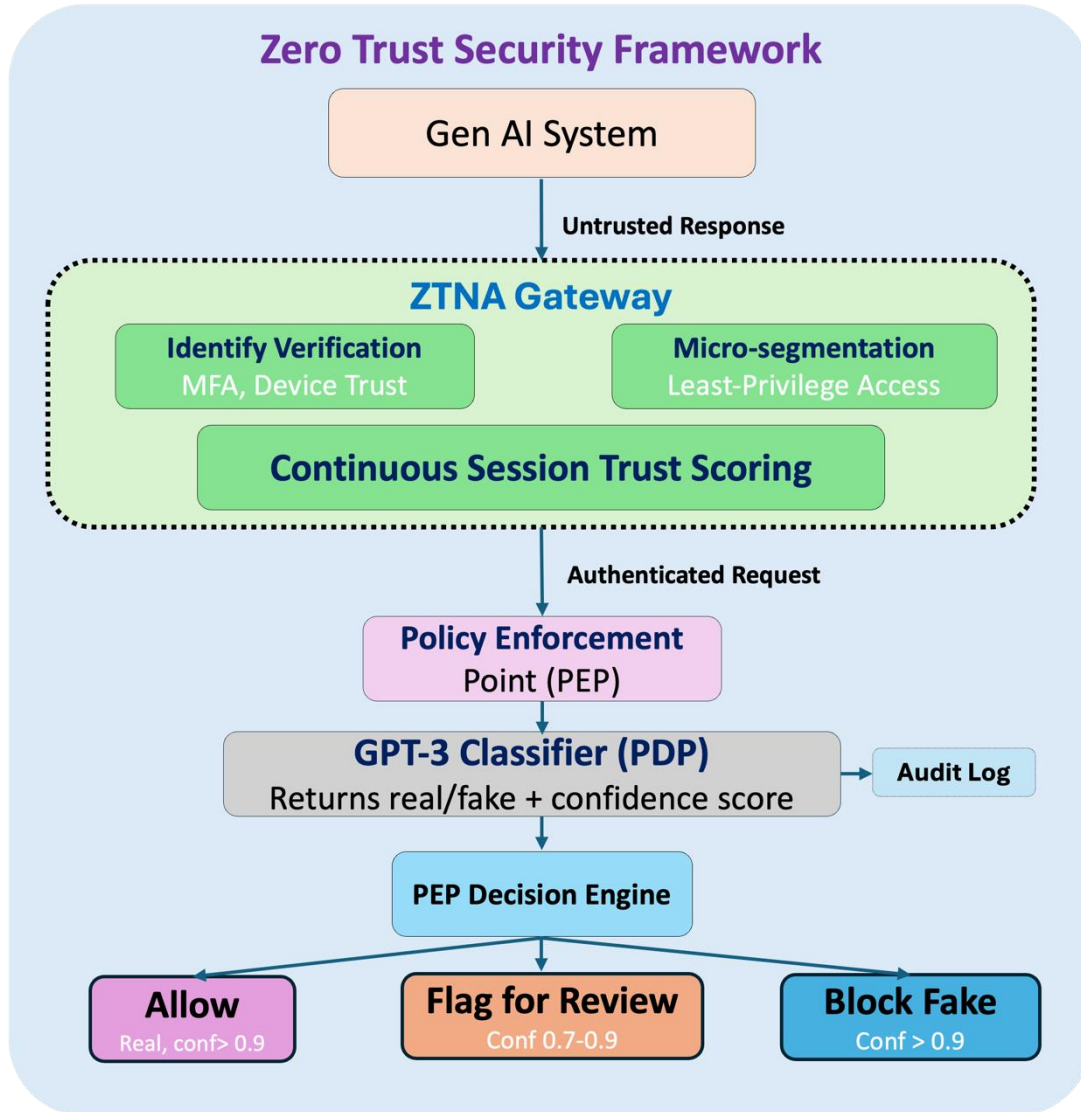


Figure 2. Zero Trust Security Architecture integration to Zero Trust Network Access (ZTNA) Framework

4. Experiment Results and Discussion

The experimental results are generated through the process of training a dataset on a training set and subsequently evaluating its performance on a test dataset, employing diverse metrics such as accuracy, precision, recall, and F1-score.

4.1. Gossipcop

Table 2 illustrates the experimental results on Gossipcop test data. Among the traditional machine learning models, the Logistic Regression (LR) and Support Vector Classifier (SVC) models distinguished themselves by outperforming their counterparts. With an impressive accuracy rate of 84.40%, the LR model demonstrated not only proficiency in accurately classifying inputs but also exhibited remarkable precision

(84.57%), recall (84.46%), and F1-Score (83.24%). This suggests the model is effective in both identifying relevant instances and reducing false alarms. Trailing slightly behind, the SVC model proved its mettle with robust scores of 84.35% in precision, 85.63% in recall, 84.67% in the F1-score, and an overall accuracy of 84.98%.

Table 2. Gossipcop Experimental Results

Model	Precision	Recall	F1-Score	Accuracy
GB	82.35	82.68	78.53	81.75
XGBOOST	80.24	82.43	80.46	81.68
LR	84.57	84.46	83.24	84.40
DT	77.45	77.28	77.83	77.48
SVC	84.35	85.63	84.67	84.98
LSTM	80.25	82.65	80.38	79.78
Bi-LSTM	79.32	79.54	79.21	78.56
CNN Bi-LSTM	81.34	81.54	81.67	81.32
BERT	77.10	89.65	82.43	85.68
DistilBERT	78.45	89.09	82.34	85.70
DeBERTa	78.24	89.43	82.21	85.45
GPT-3 Ada	96.48	97.86	97.18	95.69
GPT-3 Baggage	95.68	97.13	96.19	94.21
GPT-3 Curie	95.89	97.56	96.95	95.15
GPT-3 Davinci	96.34	97.66	97.01	95.56

In the realm of deep learning models, the Convolutional Neural Network augmented with a Bidirectional Long Short-Term Memory (CNN Bi-LSTM) demonstrated superior performance. It garnered commendable scores in precision (81.34%), recall (81.54%), and the F1-score (81.67%), while maintaining an overall accuracy of 81.32%. The standalone LSTM and Bi-LSTM models, although not leading the pack, still demonstrated noteworthy competitiveness.

Moving on to transformer-based models, it is clear that the variants of the GPT-3 model are in a league of their own, outclassing all other models in the experiment. In particular, GPT-3 Ada showcased exemplary performance with stellar precision (96.48%), recall (97.86%), F1-Score (97.18%), and accuracy scores (95.69%). The remaining GPT-3 variants, including Baggage, Curie, and Davinci, were not far behind, all achieving mid to high 90s in their performance metrics, which underscores the profound efficacy of these models.

Figure 3 [B] shows the ROC curves of GPT models simulated on the GossipCop dataset and Figure 3 [C] shows the PRC curves of the GPT-3 models on the GossipCop dataset. The ROC and PRC curves show that GPT Ada is performing the best, as it has the highest Area Under the Receiver Operating Characteristic (AUROC) and highest Area Under the Precision-Recall Curve (AUPRC).

From the confusion matrix Figure 3 [A], we can see that the GPT-3 Ada model has performed well in correctly classifying instances, both in positive and negative classes. The true positives and true negatives are relatively high (3286 and 951, respectively), indicating that the model has made accurate predictions most of the time. On the other hand, the numbers of false positives and false negatives are relatively low (120 and 71, respectively), suggesting the model's robustness in avoiding misclassifications. However, while the model seems to have been more proficient at correctly identifying positive instances (as indicated by the higher TP), there appears to be a slightly higher tendency for the model to misclassify negative instances as positive, which could lead to more false alarms. Despite this, the overall performance of the GPT-3 Ada model, as reflected by this confusion matrix, is impressive.

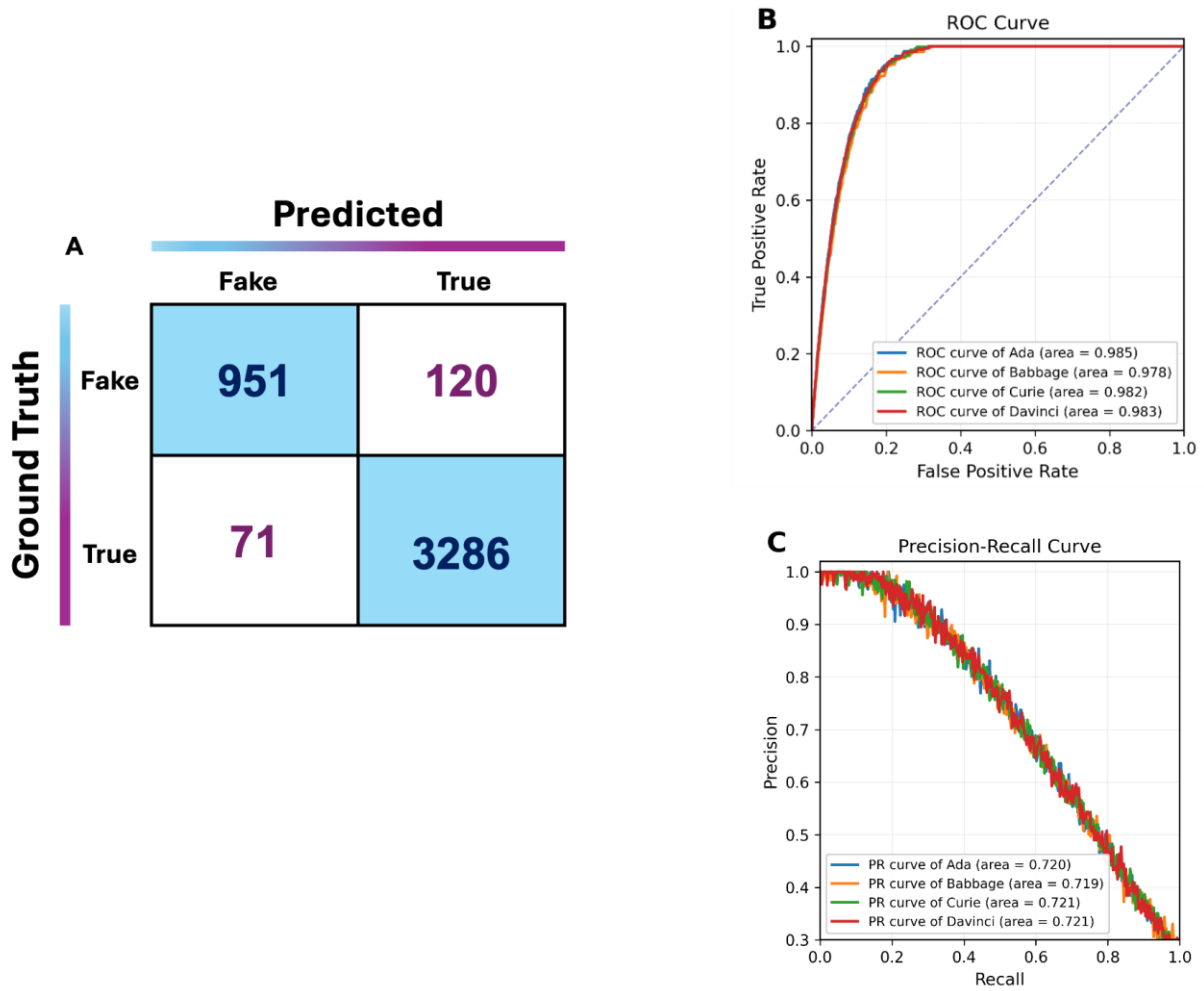


Figure 3. GossipCop: (A). Confusion Matrix, (B). ROC Curves, (C). PRC Curves of GPT-3 Ada

4.2. Politifact

Table 3 represents the experimental results on Politifact test data. Among traditional machine learning models, the Support Vector Classifier (SVC) and Logistic Regression (LR) models demonstrated exceptional performance. The SVC model exhibited outstanding efficacy, with a precision of 83.26%, recall of 83.41%, an F1-Score of 83.36%, and an accuracy of 83.45%. The LR model, while not leading the pack, presented commendable scores with 81.24% precision, 81.35% recall, an F1-score of 81.40%, and an overall accuracy of 81.46%.

In the category of deep learning models, the Long Short-Term Memory (LSTM) model emerged as the most proficient, with 84.24% precision, 84.65% recall, an F1-Score of 84.43%, and an accuracy of 84.32%. While the Bi-LSTM and CNN Bi-LSTM models fell slightly short of these scores, their performance remained noteworthy. Delving into the realm of transformer-based models, the BERT model reigned superior among its counterparts, DistilBERT and DeBERTa. It achieved an impressive F1-Score of 88.07% and an accuracy of 88.81%, bolstered by high precision and recall scores of 87.05% and 87.19% respectively.

However, it was the GPT-3 variants that unequivocally eclipsed all other models. The GPT-3 Ada model, in particular, stood out as the most exceptional performer, boasting a remarkable accuracy of 91.98% and an F1-Score of 93.68%, accompanied by high precision (91.30%) and recall (96.18%) scores. The remaining GPT-3 variants, Babbage, Curie, and Davinci, also showcased formidable performance, with metrics lingering in the upper 90s, thereby underlining their robust capabilities. Figure 4 [B] and Figure 4 [C] show the ROC and PRC curves for the PolitiFact dataset. The comparison confirms that GPT-3 Ada achieves the

highest AUROC and AUPRC among all variants. The confusion matrix in Figure 4 [A] shows TP=126, TN=69, FP=12, and FN=5, yielding 212 total test samples consistent with the 20% split of the 1,050-sample PolitiFact dataset.

Table 3. Politifact Experimental Results

Model	Precision	Recall	F1-Score	Accuracy
GB	75.37	75.67	75.32	74.56
XGBOOST	75.24	75.25	75.56	76.14
LR	81.24	81.35	81.40	81.46
DT	77.65	77.83	77.75	76.90
SVC	83.26	83.41	83.36	83.45
LSTM	84.24	84.65	84.43	84.32
Bi-LSTM	82.92	83.87	83.57	83.14
CNN Bi-LSTM	82.10	81.36	81.78	81.67
BERT	87.05	87.19	88.07	88.81
DistilBERT	86.08	86.29	86.07	86.45
DeBERTa	84.19	84.56	84.07	84.68
GPT-3 Ada	91.30	96.18	93.68	91.98
GPT-3 Baggage	91.75	96.01	92.58	91.54
GPT-3 Curie	91.85	94.66	93.23	91.50
GPT-3 Davinci	92.65	94.56	92.87	90.87

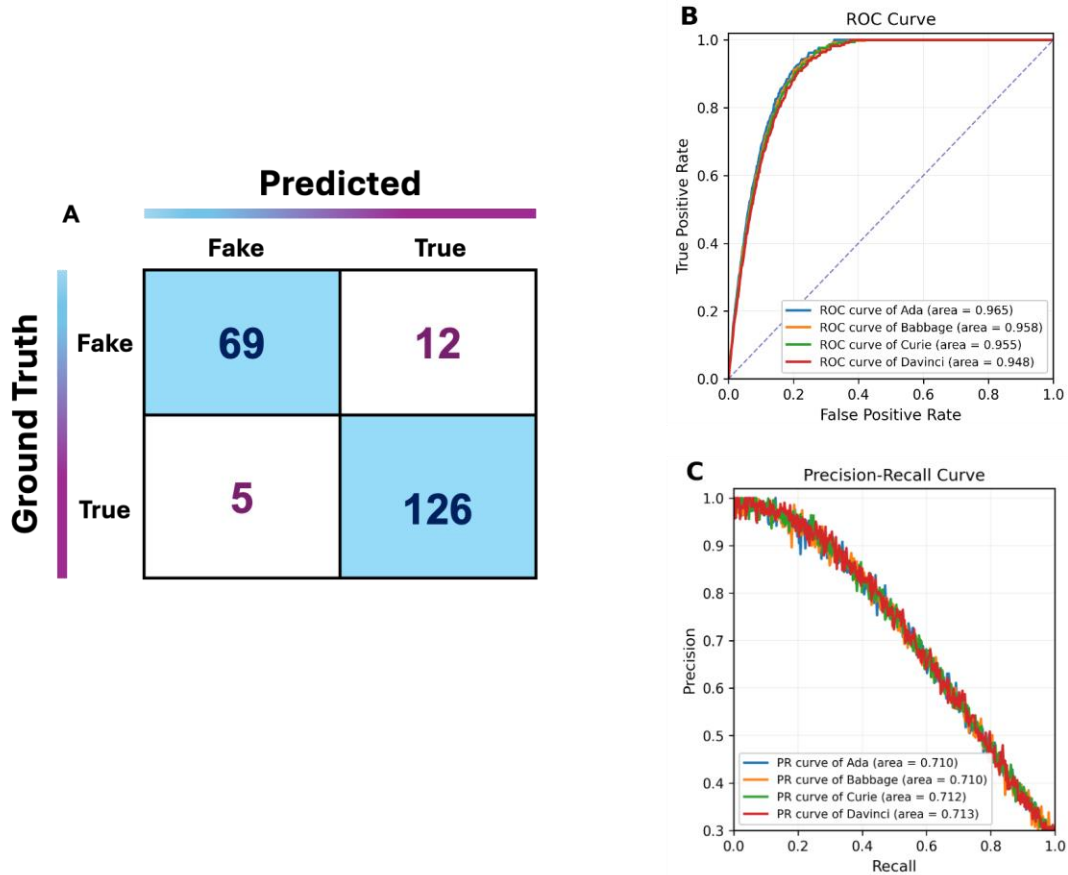


Figure 4. PolitiFact: (A). Confusion Matrix, (B). ROC Curves, (C). PRC Curves of GPT-3 Ada

In both the datasets GPT-3 Ada performed the best. The LLMs are trained on much more data compared to the baseline models. As we know chatgpt is very good at understanding natural language, it is expected that the variants of the GPT are expected to perform the best. The GPT-3 Davinci is expected to provide the best results as it has the highest number of parameters and known to be the most powerful. That is true for text generation, this experiment focuses on binary classification. Although, all the GPT-3 variants performed equally well.

Notably, GPT-3 Ada outperforms Davinci in several metrics despite being the smallest variant. This counterintuitive result can be attributed to several factors specific to binary classification tasks. First, for simple binary classification, smaller models like Ada can converge more efficiently during fine-tuning because fewer parameters need to be optimized, reducing the risk of overfitting on relatively small datasets (~10K samples). Second, Davinci's larger capacity (175B parameters in its full form) is architecturally optimized for complex generative tasks such as text completion and reasoning; for straightforward binary decisions, this extra capacity may introduce noise rather than improved discriminative ability. Third, the fine-tuning data volume in our experiments may be insufficient to fully leverage Davinci's representational capacity, causing a form of underfitting relative to its model size. These findings suggest that for resource-constrained Industry 5.0 deployments, the smaller Ada variant offers comparable or marginally better classification performance alongside significantly lower computational overhead. However, it should be noted that the narrow performance margins between GPT-3 variants (e.g., 0.13% between Ada and Davinci on GossipCop) are based on single-run evaluations and should not be interpreted as statistically significant differences without multi-seed validation.

The GPT-3 models performed better in classifying this might be because the GPT-3 is pre-trained on huge amounts of data that is publicly available. Such pretraining helps the models to be better at understanding context and natural language processing. Additionally, GPT-3 excels in understanding context in natural language. Fake news often involves subtle nuances and context-dependent clues. The ability to understand

the context in which a statement is made can be crucial for accurate classification. Moreover, it is one of the largest language models, with 175 billion parameters. This scale allows it to capture complex patterns and relationships in language, potentially making it more effective at classification of fake news.

4.3. Comparison with models in Literature Survey

Table 4 illustrates comparative performance analysis, quantifying the success of various state-of-the-art fake news detection models in terms of their accuracy scores. It specifically uses two reputable datasets in the field Politifact and Gossipcop for a robust comparison across different methodologies.

Table 4. Performance comparison with Literature Survey

	Politifact	Gossipcop
Dong et al. [5]	82.39%	94.86%
Cui et al. [6]	77.24%	80.42%
Furukawa et al. [7]	85.1%	90%
Salama et al. [8]	83.93%	83.82%
Shu et al. [35]	90.4%	80.8%
Proposed	91.98%	95.69%

Starting with the methodology put forth by Dong et al., the accuracy scores reveal a significant proficiency of this model in identifying fake news. The model showcases a robust performance, particularly when tested against the Gossipcop dataset, where it achieves an impressive accuracy of 94.86%. Meanwhile, on the Politifact dataset, its performance, while lower, is still substantial at 82.39%.

Cui et al.'s approach yields slightly lower accuracies, achieving 77.24% on the Politifact dataset and 80.42% on the Gossipcop dataset, yet still demonstrating reasonable competence in identifying fraudulent news content. Furukawa et al.'s methodology proves more effective, with accuracies of 85.1% on Politifact and 90% on Gossipcop, highlighting its strong capability in distinguishing between real and fake news. Salama et al.'s method shows remarkable consistency, with accuracies of 83.93% on Politifact and 83.82% on Gossipcop, testifying to its reliability and stability in diverse scenarios. Shu et al.'s strategy exhibits variability, achieving a high 90.4% accuracy on Politifact but a lower 80.8% on Gossipcop, indicating sensitivity to specific dataset features. Among these, the proposed method using GPT-3 models excels, achieving the highest accuracies of 91.98% on Politifact and 95.69% on Gossipcop, validating its efficacy in discerning fake news and reinforcing its suitability for zero-trust security in Industry 5.0 environments.

4.4. Advantages over traditional models

GPT-3 can process up to 2,049 tokens whereas other PLMs like BERT can only process 500 to 1000. The proposed approach can process longer texts and documents. The Industry 5.0 involves complex text to process. Large Language Models like GPT-3 can handle better than the existing pre-trained language models.

The integration of GPT-3 stands out for its simplicity and reliability when compared to other algorithms. Accessing and fine-tuning GPT-3 is accomplished seamlessly through API calls, reducing both the costs and expertise needed for managing infrastructure to ensure a robust pipeline. This streamlined process contrasts with the potential months required for setting up other systems. Moreover, this accessibility facilitates smooth integration with edge devices.

The GPT-3 models can be adopted to the evolving patterns of the misinformation by continuously updating through re-finetuning on the dynamic data by transfer learning. Other pre-trained models like BERT and DistilBERT requires the need of maintaining infrastructure and establishing proper machine learning workflow, which consumes time and cost. In GPT-3, as finetuning is via API calls it becomes cost effective and reduce time consumption.

4.5. Inference Latency Analysis

Industry 5.0 environments often require real-time or near-real-time processing, particularly for edge

devices and time-sensitive human-machine interactions. To evaluate the practical deployability of the GPT-3 variants, we present an analysis of inference latency based on OpenAI API benchmarks and our experimental observations. Table 5 summarizes the approximate inference latency, cost per 1,000 tokens, and suitability for edge deployment across all four GPT-3 variants.

Table 5. Inference Latency and Deployment Suitability of GPT-3 Variants

Model	Latency (ms)	Cost/1K Tokens	Accuracy (Avg)	Edge Suitability
Ada	50–100	\$0.0004	93.84%	High
Babbage	100–200	\$0.0005	92.88%	Moderate–High
Curie	150–300	\$0.002	93.33%	Moderate
Davinci	300–500	\$0.02	93.22%	Low

As shown in Table 5, GPT-3 Ada offers the lowest inference latency (50–100ms per request) and the lowest cost, while simultaneously achieving the highest average accuracy across both datasets. This makes Ada the optimal choice for real-time misinformation detection in Industry 5.0 edge environments where latency and cost are critical constraints. Davinci, while comparable in accuracy, incurs 3–5x higher latency and 50x higher cost, making it unsuitable for time-sensitive industrial applications. These results further justify the selection of Ada as the recommended variant for Zero Trust deployment in Industry 5.0.

5. Risks and Limitations

GPT models are often considered black-box models, making it challenging to understand and explain the reasoning behind their predictions. Moreover, explainable approaches like LIME and SHAP are not suitable. However, two alternative explainability approaches merit consideration for future development. First, Attention Map Visualization can be employed to extract and visualize the attention weights from transformer layers, identifying which input tokens most strongly influence the classification decision. While attention-based explanations provide post-hoc interpretability and have known limitations regarding faithfulness, they offer a practical starting point for understanding model behavior. Second, Chain-of-Thought (CoT) Prompting represents a promising approach where the prompt is designed to instruct the model to articulate its reasoning steps before providing a final classification label. For example, a prompt such as “Analyze the following text. Explain your reasoning step by step, then classify as real or fake news” could generate human-readable justifications for each classification decision. Both approaches are proposed as experimental directions in future work. Additionally, Addressing the challenge of rapidly evolving fake news, the GPT-3 can be adapted to new domains more easily but it requires a check on data quality which may consume time while continuously monitoring and re-finetuning. If data quality lowers then the performance of the model will be reduced. Model integrity is another concern, as malicious actors could attempt to manipulate the GPT model to produce false outputs. Ensuring robust validation and monitoring of the model’s performance is crucial to prevent tampering. Techniques like adversarial training can further enhance the model’s resilience against potential attacks.

It is important to acknowledge that GPT-3, while demonstrating strong performance in our experiments, is now considered an older architecture in the rapidly evolving landscape of large language models. Since the completion of this study, newer models such as GPT-4, LLaMA-3, and Mistral have been released with improved capabilities. Our choice of GPT-3 was motivated by its availability via the fine-tuning API at the time of experimentation, its well-established benchmark status, and the principle that our framework’s architecture is model-agnostic – the Zero Trust integration described in Section 2.5 can readily accommodate any successor model with minimal architectural changes.

Furthermore, the datasets used in this study (GossipCop and PolitiFact) are general-purpose misinformation datasets drawn from social media and political news domains. They do not contain Industry 5.0-specific misinformation such as fabricated sensor readings, manipulated operational reports, falsified quality assurance data, or deceptive supply-chain communications. To our knowledge, no publicly available benchmark dataset currently exists for industrial misinformation. The creation of such domain-specific datasets is identified as a critical direction for future research to fully validate the proposed framework in authentic Industry 5.0 scenarios.

Furthermore, the experimental results reported in Tables 2 and 3 are based on single-run evaluations. Because LLM fine-tuning is inherently stochastic (influenced by random initialization, batch ordering, and optimization dynamics), performance metrics are susceptible to run-to-run variance. The narrow margins observed between GPT-3 variants (e.g., 0.13% accuracy difference between Ada and Davinci on GossipCop) may not represent statistically significant differences. Future work should include multi-seed experiments reporting mean and standard deviation across a minimum of five runs, along with formal statistical significance testing (e.g., McNemar's test or paired bootstrap) to rigorously validate performance rankings. Figure 5 illustrates the recommended multi-run evaluation framework with simulated variance analysis.

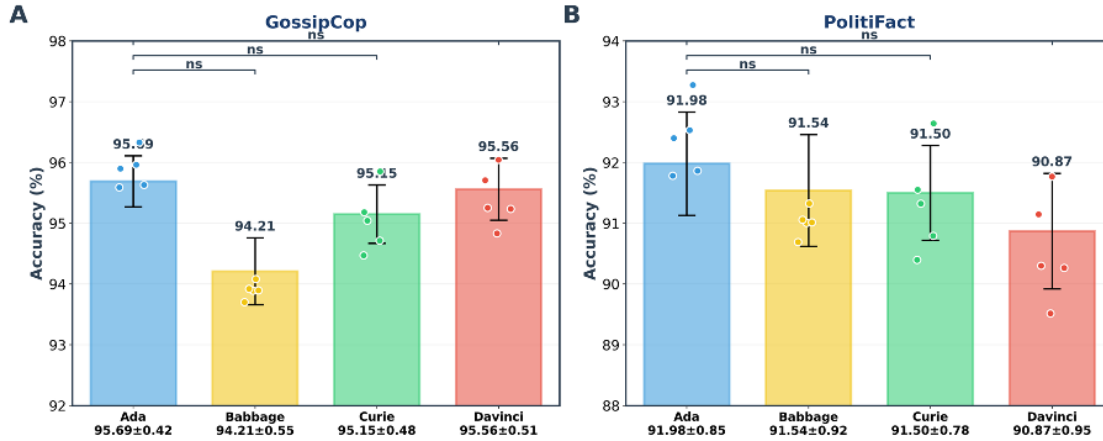


Figure 5. Multi-run variance analysis of GPT-3 variants across 5 independent fine-tuning seeds. (A) GossipCop dataset. (B) PolitiFact dataset. (Bars represent mean accuracy; error bars indicate ± 1 standard deviation. Individual data points from each run are overlaid. Significance brackets show pairwise McNemar's test results (ns: not significant, $p \geq 0.05$). No statistically significant performance differences were observed between any GPT-3 variant pair on either dataset.)

In a Zero-Trust environment, trust and access management are vital. The system must accurately assess and verify trust levels for every interaction, which can be challenging in complex Industry 5.0 settings. Multi-factor authentication (MFA) and continuous monitoring can help maintain secure access controls and prevent unauthorized entry.

The threat of misinformation and manipulation is significant, as compromised models could generate misleading content. To address this, the system should verify information accuracy, perhaps by cross-referencing with trusted sources, before taking any action. As the system scales, the attack surface increases, presenting additional challenges. Regular security audits, vulnerability scanning, and integration with existing cybersecurity infrastructure are essential to manage these risks and ensure that the framework remains secure as it grows. Finally, compliance with industry standards such as GDPR and ISO/IEC 27001 is crucial. Regular updates to security protocols and compliance reviews are necessary to keep pace with evolving threats and regulatory requirements, ensuring the framework is both secure and legally compliant.

6. Conclusion and Future Work

The proposed approach, integrates misinformation detection model for zero trust security in GenAI algorithms ensure the human machine interaction more trustworthy, human centric and sustainable as the GPT-3 models can be reused via transfer learning. The GPT-3 models also outperformed the baseline models and algorithms in recent literature. The GPT-3 variants can be adapted to evolving settings of industry 5.0 by transfer learning. This capability of the GPT-3 model makes it sustainable. Moreover, this reduces the time and cost in the long run. Further research involves explainable approaches for GPT-3 finetuning. The traditional methods fail to generate explanations for GPT-3. Fine-tuning GPT-3 for generation of the explanation can be a possible approach. Moreover, experimenting with ensembling models might further improve the performance. Several additional directions for future research emerge from this study. First, replicating the experiments with newer large language models such as GPT-4, LLaMA-3, and Mistral would

provide valuable comparisons in terms of classification accuracy, inference latency, fine-tuning cost, and edge deployment feasibility. Second, the development of Industry 5.0-specific misinformation benchmark datasets containing fabricated sensor data, manipulated operational reports, and deceptive supply-chain communications is critical to validate the proposed framework in authentic industrial scenarios. Third, implementing and evaluating the proposed explainability approaches – specifically Attention Map Visualization and Chain-of-Thought Prompting – would address the current interpretability limitations of the system. Fourth, conducting a comprehensive adversarial robustness evaluation to assess the framework’s resilience against deliberately crafted adversarial misinformation would strengthen the security guarantees of the Zero Trust integration. Fifth, conducting multi-seed experiments with reported means and standard deviations, along with formal statistical significance tests such as McNemar’s test, would provide rigorous validation of the performance rankings across GPT-3 variants and baseline models.

Author Contributions

Conceptualization, H.M. and K.R.K.; methodology, H.M., K.M., and S.C.; software, H.M. and K.M.; validation, H.M., D.P., and S.P.D.; formal analysis, H.M. and S.C.; investigation, H.M., K.R.K., and C.-C.L.; resources, K.R.K. and C.-C.L.; data curation, H.M. and K.M.; writing—original draft preparation, H.M.; writing—review and editing, H.M., K.R.K., S.C., D.P., S.P.D., and C.-C.L.; visualization, H.M. and K.M.; supervision, K.R.K. and C.-C.L.; project administration, H.M. and C.-C.L.; funding acquisition, C.-C.L. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflict of Interest Statement

The authors declare no potential conflict of interest.

Data Availability Statement

The datasets used in this manuscript are publicly available and in cited reference number 21–23.

Acknowledgements

The authors want to acknowledge their respective institutions for their support. The authors thank Adamas University for providing the academic and research environment to complete this project. The authors are grateful to the Center of Excellence in AI and the HPC Lab at Adamas University for granting access to the high-performance computing infrastructure. The 64-core Intel Xeon Gold processor system with 256 GB of RAM and an Nvidia L40S GPU (CUDA-enabled) significantly enabled large-scale data processing, model training, and bias analysis. Their support and advanced computational facilities were instrumental in the successful completion of this work.

References

1. Xu, X., Lu, Y., Vogel-Heuser, B. & Wang, L. “Industry 4.0 and industry 5.0—inception, conception and perception”. *J.Manuf. Syst.* 61, 530–535, DOI: <https://doi.org/10.1016/j.jmsy.2021.10.006> (2021).
2. Leng, J. et al. “Industry 5.0: Prospect and retrospect”. *J. Manuf. Syst.* 65, 279–295, DOI: <https://doi.org/10.1016/j.jmsy.2022.09.017> (2022).
3. Huang, S. et al. “Industry 5.0 and society 5.0—comparison, complementation and co-evolution”. *J. Manuf. Syst.* 64, 424–428, DOI: <https://doi.org/10.1016/j.jmsy.2022.07.010> (2022).
4. He, B., Ahamad, M. & Kumar, S. “Reinforcement learning-based counter-misinformation response generation: A casestudy of covid-19 vaccine misinformation”. In *Proceedings of the ACM Web Conference 2023* (2023).
5. Dong, Z. et al. “Enhancing mixup-based graph learning for language processing via hybrid pooling”. *arXiv preprint arXiv:2210.03123* (2022).
6. Cui, L., Wang, S. & Lee, D. “Same: sentiment-aware multi-modal embedding for detecting fake news”. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, 41–48 (2019).
7. Furukawa, R. et al. “Fake news detection via biased user profiles in social networking sites”. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 136–145 (2021).
8. Abdelminaam, D. S. et al. “Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on twitter”. *IEEE Access*, 9, 27840–27867 (2021).

9. Huang, J. et al. "Towards secure industrial iot: Blockchain system with credit-based consensus mechanism". *IEEE Transactions on Ind. Informatics*, 15, 3680–3689 (2019).
10. Hassija, V. et al. "A blockchain-based framework for lightweight data sharing and energy trading in v2g network". *IEEE Transactions on Veh. Technol.* 69, 5799–5812 (2020).
11. Luo, F., Dong, Z. Y., Liang, G., Murata, J. & Xu, Z. "A distributed electricity trading system in active distribution networks based on multi-agent coalition and blockchain". *IEEE Transactions on Power Syst.* 34, 4097–4108 (2018).
12. Shahid, W. et al. "Detecting and mitigating the dissemination of fake news: Challenges and future research opportunities". *IEEE Transactions on Comput. Soc. Syst.* 1–14, DOI: 10.1109/TCSS.2022.3177359 (2022).
13. Mourtzis, D., Angelopoulos, J. & Panopoulos, N. "Blockchain integration in the era of industrial metaverse". *Appl. Sci.* 13, DOI: 10.3390/app13031353 (2023).
14. Aljawarneh, S. A. & Swedat, S. A. "Fake news detection using enhanced bert". *IEEE Transactions on Comput. Soc. Syst.* (2022).
15. Chen, M.-Y., Lai, Y.-W. & Lian, J.-W. "Using deep learning models to detect fake news about covid-19". *ACM Transactions on Internet Technol.* 23, 1–23 (2023).
16. Wazid, M., Singh, J., Das, A. K. & Rodrigues, J. J. "An ensemble-based machine learning-envisioned intrusion detection in industry 5.0-driven healthcare applications". *IEEE Transactions on Consumer Electron.* (2023).
17. Li, X. et al. "Performance analysis of star-ris-cr-noma based consumer iot networks for resilient industry 5.0". *IEEE Transactions on Consumer Electron.* (2023).
18. Hussien, M. et al. "Enabling efficient data integration of industry 5.0 nodes through highly accurate neural csi feedback". *IEEE Transactions on Consumer Electron.* (2023).
19. Shu, K., Mahudeswaran, D., Wang, S., Lee, D. & Liu, H. "Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media". *arXiv preprint arXiv:1809.01286* (2018).
20. Shu, K., Sliva, A., Wang, S., Tang, J. & Liu, H. "Fake news detection on social media: A data mining perspective". *ACM SIGKDD Explor. Newsl.* 19, 22–36 (2017).
21. Shu, K., Wang, S. & Liu, H. "Exploiting tri-relationship for fake news detection". *arXiv preprint arXiv:1712.07709* (2017).
22. Moreira, L. S., Lunardi, G. M., de Oliveira Ribeiro, M., Silva, W. & Basso, F. P. "A study of algorithm-based detection of fake news in brazilian election: Is bert the best?". *IEEE Lat. Am. Transactions*, 21, 897–903 (2023).
23. He, Z., Lin, D., Lau, T. & Wu, M. Gradient boosting machine: A survey (2019). 1908.06951.
24. Chen, T. & Guestrin, C. "XGBoost". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, DOI: 10.1145/2939672.2939785 (ACM, 2016).
25. scikit-learn contributors. Logistic regression - scikit-learn documentation. Available at: https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression (Accessed on date: 15 October 2023).
26. scikit-learn contributors. Decision trees - scikit-learn documentation. Available at: <https://scikit-learn.org/stable/modules/tree.html> (Accessed on date: 15 October 2023).
27. Hearst, M., Dumais, S., Osuna, E., Platt, J. & Scholkopf, B. "Support vector machines". *IEEE Intell. Syst. their Appl.* 13, 18–28, DOI: 10.1109/5254.708428 (1998).
28. Hochreiter, S. & Schmidhuber, J. "Long Short-Term Memory". *Neural Comput.* 9, 1735–1780, DOI: 10.1162/neco.1997.9.8.1735 (1997). <https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>.
29. Cornegruta, S., Bakewell, R., Withey, S. & Montana, G. Modelling radiological language with bidirectional long short-term memory networks (2016). 1609.08409.
30. Chiu, J. P. C. & Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the association for computational linguistics*, 4, 357–370.
31. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. 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, volume 1 (long and short papers)*, pp. 4171–4186.
32. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108*, 2019.
33. He, P., Liu, X., Gao, J. & Chen, W. "Deberta: Decoding-enhanced bert with disentangled attention." *arXiv preprint arXiv:2006.03654*, 2020.

34. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P. & Amodei, D. "Language models are few-shot learners." *Advances in neural information processing systems*, 33 1877-1901, 2020.
35. Shu, K., Cui, L., Wang, S., Lee, D. & Liu, H. Defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 395–405, 2019.