Grenze International Journal of Engineering and Technology, June Issue

A BERT-based Ethical Framework for Real-Time Toxicity Detection and Moderation in Social Media

Shabana Sultana[1] and Harshitha H S[2]
[1-2]The National Institute of Engineering, Mysuru-570008, Karnataka, India
Email: shabana@nie.ac.in, harshithanayaka25@gmail.com

***Abstract*— The transformer model is a crucial part of Natural Language Processing. Among all the transformers, Bidirectional Encoder Representations from Transformers, otherwise known as BERT, is one of the models that have greatly impacted the way computers are able to understand and process text information. The traditional model used to understand Natural Language is often faced with difficulties in trying to obtain a deep understanding of the information provided, as it only does so in one way. For example, in processing a sentence that goes from left to right. In the bidirectional model of BERT, it is able to process in both ways, which are right to left and left to right.**

BERT is primarily used in this research work for the purpose of understanding the text as well as obtaining relevant information from a large number of documents. BERT can be used for its capacity to provide a deep understanding of words, based on the context in which the words are used.

The results achieved from the experiment have proved that the transformer models, like BERT, are much more superior to the rule-based systems and statistical approaches. The most important benefit that BERT provides is its efficiency in handling deeper contexts by providing information more easily and quickly as it handles deeper contexts in almost all NLP tasks.

***Keywords*— BERT, Transformer, NLP, Contextual Encoding, Semantic Search, Dense Retrieval, Text Classification, Question Answering, Intent Detection.**

I. INTRODUCTION

The complexity of natural language transcends the limitations of early NLP systems, which were designed and developed by means of relatively simple statistical models. The development of transformer models and generative AI technology helped to greatly enhance the processing of meaningful texts by NLP systems [11]. A paradigm shift was caused by the appearance of the Transformer model, which replaced traditional recurrent structures in neural networks using self-attention and was able to handle long-term accuracy in texts much more successfully. Among the models of the type of transformers, BERT should be noted as one of the most popular ones, thanks to its property of processing texts in both directions [7][11]. BERT is not developed for generating texts, but its bidirectionally encoding enables it to gain in-depth comprehension of texts, which makes it extremely successful in information retrieval, text categorization, or semantic searching [9].

In RAG systems, BERT is frequently used as a retriever, providing the generative model with accurate and contextually rich information to reduce factual errors and hallucinated outputs [4][9][10]. This work presents a Twitter-like social media platform through employing BERT in three major modules.

Grenze ID: 01.GIJET.12.2.245_26
© Grenze Scientific Society, 2026

The first module enables content creation by providing the user with relevant and context-aware text suggestions. The second module gives the toxicity recognition, within user-generated messages, detecting malicious or offensive language [3][6]. The third module is an analytics dashboard, which showcases real-time insights into user activity and the health of the platform.

One of the main aims of this framework is facilitating ethical and transparent system processes. Rather than incorporating responsible AI elements as an additional consideration, ethics are implemented directly in the process. The intention is to prove how language models can be used securely, fairly, and reliably while also promoting positive interactions on the internet[10].

This paper gives an ethical moderation framework for social media platforms that leverages BERT to enhance content generation, toxicity detection, and semantic retrieval. The major contributions of this paper are as follows:

- A real-time contextual content assistance module using BERT for responsible content generation.
- An optimized BERT model for toxicity detection that performs better than traditional LSTM and rule-based models.
- A dense semantic retrieval module using BERT embeddings for retrieval-augmented content generation.
- An ethical moderation pipeline that tracks user behavior and analytics for fair, transparent, and scalable online content management.

II. Related Work

There have been various studies that prove the ever-increasing efficiency of transformer models in the identification of toxic/hateful posts in social networks. In particular, numerous experiments carried out on platforms like Twitter and Instagram have established that language models are more efficient compared to traditional ML approaches like SVM or Naïve Bayes classifiers because they can determine not just the meaning of the expression used, but the semantics of the word as part of the broader context in which it was made. This feature is especially crucial for social media platforms where toxic content may be hidden by contextuality and sarcasm.

One of the most prominent trends in recent years is associated with the design of adaptive moderation systems capable of dealing with the rapid evolution of online language. Traditional approaches that involve the identification of fixed keywords frequently overlook new slurs, coded language, or obfuscation tactics adopted by users to disguise their messages. To overcome this challenge, several scientists tried using BERT along with dynamic lexicon expansion and word-embedding vocabulary augmentation. Both approaches proved more efficient at capturing unknown toxic terms without significantly lowering accuracy rates. Another strategy employed by researchers has involved vocabulary augmentation.

The other relevant avenue pertains to designing computation-friendly moderation models. While transformer architectures exhibit impressive performance, deploying them in live applications is not easy due to computational considerations. To overcome this obstacle, recent work has experimented with low-parameter fine-tuning of these networks through parameter-efficient methods like LoRA that decrease the number of trainable parameters significantly without compromising the detection accuracy. This issue becomes more pertinent in large social networking sites where rapid detection of abusive content is essential.

However, there are still some problems that arise from them. The existing model can still have social biases in their training data, lack explainability, and fail in recognizing toxicities, sarcasms, and other new types of offensive content in specific contexts.

In addition, most researches concentrate solely on improving the accuracy of toxicity recognition without considering the issues related to ethical moderation and transparency in their work. This issue implies that the future research should address the need for toxicity detection systems which are able to incorporate ethics in their decision-making process.

III. Novelty and Contributions

Current approaches for moderating social media mainly concentrate on standalone systems for toxicity detection, where any posted text is classified into a particular class. Current BERT-based models mostly offer single-task moderation without a comprehensive system that covers content creation, toxicity detection, behavior analysis, and moderation in real-time. Also, most of the current tools depend on classification score alone and may not identify semantically similar abusive content or misuse.

A disadvantage in using explainability models like SHAP and LIME for identifying toxic posts in social media is increased computation costs due to explainability models. Consequently, there is a need for a lightweight yet scalable solution for ensuring user safety while maximizing engagement in social media platforms.
The present work proposes a new approach to solving the above problem by designing a multi-modal BERT model that covers content creation, toxicity detection, behavior analysis, and moderation in real-time.

IV. METHODOLOGY

In the proposed work, a social platform like Twitter is developed, which aims for improved engagement with the user, at the same time facilitating the moderation of the content in a responsible manner. This is achieved through the implementation of the Advance Natural Language Processing technique, which is developed by using the BERT algorithm, and helps the user in preparing engaging content and moderating the content shared by the users automatically. In this context, it is again an effort for the improvement of the users and providing them with a better experience in the online world.
It comprises of three major modules. The Content Creation Module, where BERT would try to help the user in thinking of a contextually relevant word or phrase towards the completion of their write-up. This is done on the basis of the text that is being put forth.
The second component is the Toxicity Detection Module. Each of these posts is then analyzed through a BERT model that has been fine-tuned on corpora of posts that include abusive or profane material. In the case of finding posts that include abusive or violent material, the posting is deemed for moderation. Accounts that abuse posting policy can be temporarily or permanently disabled.
The third module is the Analytics and Dashboard module, which gives a real-time report on the activity that is taking place in the platform. This module presents statistics like the number of active users, the number of accounts that have been reported for abusive acts, and the number of posted messages deemed to be in violation of the content guidelines.
The preprocessing pipeline includes text normalization, tokenization, and BERT-based embedding to prepare user input for both generation and classification tasks. The toxicity detection module takes advantage of attention-weighted embeddings in addition to using cosine similarity to detect harmful material. Once a certain threshold is surpassed by the concentration of toxic expressions, actions in regard to moderation are triggered to ensure that community guidelines are maintained. Figure 1 presents an overview of the system design. Because analysis in real-time is needed, SHAP or LIME is not used as an explanation technique. Instead, interpretation is achieved through monitoring.
To test the effectiveness of the framework, both synthetic and publicly accessible social media data was employed. The toxicity detection module was trained on a combination of publicly available benchmark datasets of toxic and hate speech tweets and synthetically generated user content curated to simulate real-world social media scenarios. The performance was measured in terms of precision, recall, and F1-score for the toxicity classification task and BLEU score for content generation tasks. The experiment shows that BERT is quite efficient in generating appropriate content and is accurate in identifying toxic content.
With the consideration of ethics of AI and UX incorporated into the design of the system, the proposed framework ensures this usefulness by providing scalability, equity, and usability with regard to their online interactions. The proposed model also portrays the use of AI-powered moderated platforms not only being useful but also applicable in the increasing demand for online communication safely through social media platforms.
To verify the effectiveness of the suggested framework, the fine-tuned BERT moderation model was benchmarked against traditional machine learning classifiers, such as SVM and LSTM, as well as state-of-the-art transformer architectures, including DistilBERT, ALBERT, RoBERTa, and zero-shot moderation based on GPT. It was found that SVM and LSTM models exhibited low contextual comprehension, especially when dealing with sarcasm, implicit toxicity, and linguistically complicated abuse messages. The DistilBERT model offered faster inference and compact model size but had slightly worse classification precision. In turn, the ALBERT model produced similar results to BERT with significantly less number of parameters; however, the recall rate for minority toxic classes was insufficient. RoBERTa achieved high performance levels and showed similarity with BERT in various metrics, but it demanded more resources and more training time than BERT. The GPT model demonstrated excellent zero-shot performance but had high latency and resource consumption, which limited its usability in real-time applications. Overall, the proposed BERT-based moderation framework proved to have the best compromise between classification accuracy, speed, computation costs, and practical application.

V. IMPLEMENTATION

The application of this project involves the use of BERT as a language model in different parts of a social media platform, which is meant for an intelligent content management process. The project is modular and includes a variety of NLP pipelines, each intended for a different job, for instance, semantic retrieval, identification of abusive contents, and assessment of quality. Below is a brief explanation of each part, along with its working equation.

BERT algorithm

- Input Representation - Tokenize text, add [CLS] and [SEP], create embeddings (token + segment + position).
- Encoding - Pass through Transformer encoders using bidirectional self-attention.
- Pre-training -
  - Masked Language Modelling (MLM).
  - Next Sentence Prediction (NSP).
- Fine-tuning - Add the task-specific layer (classification, QA, NER, etc.) and train end-to-end.
- Prediction - Use [CLS] or token embeddings for final output.

A. BERT Input Representation

To create embeddings from user input, each input is broken down into tokens and embedded according to the input form of the BERT model. The resulting representation is created by combining embeddings of tokens, segments, and positions.

$$\text{Input} = E_{\text{token}} + E_{\text{segment}} + E_{\text{position}} \quad \text{----------(1)}$$

This allows the model to maintain syntactic order and differentiate between segments.

B. Self-Attention and Encoding

BERT uses self-attention to handle dependencies between the entire sentence. Every word in BERT attends to all other words based on the attention score computed from the query, key, and value vectors. The attention modules in BERT allow it to capture context about every word.

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{----------(2)}$$

.

This attention output is passed through multiple transformer layers to generate final contextual embeddings used in downstream tasks.

C. Offensive Content Detection

For classifying posts as offensive or non-offensive, the [CLS] token output from BERT is passed through a fully connected layer followed by a softmax activation. This provides probability scores for each class label (e.g., toxic, non-toxic).

$$\hat{y} = \text{softmax}(W \cdot h_{[\text{CLS}]} + b) \quad \text{----------(3)}$$

D. Classification Training Objective

The system uses the cross-entropy loss function to optimize the classification task. This function measures the error between predicted probability and class true labels during training.

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad \text{----------(4)}$$

By minimizing this loss, the model becomes better at distinguishing between harmful and safe content.

E. Semantic Retrieval Using BERT Embeddings

In order to increase content discovery and enhance context matching, the use of word embeddings based on BERT has been added in this context. In this process, every sentence will result in a dense representation, and the similarity between the user query, as well as the existing posts, will be calculated based on cosine similarity.

$$\text{Cosine Similarity}\,(A,B) = \frac{A \cdot B}{\|A\|\|B\|} \quad \text{----------(5)}$$

This mechanism is essential for fetching relevant content or evidence before generation or moderation.

F. Content Quality Evaluation (ROUGE-L Score)

To check the quality of the generated text, we use the ROUGE-L metric. This metric evaluates the longest common subsequence between the system's output and a reference sentence, giving an indication of how much useful information is captured and how good the response is.

$$\text{ROUGE-L} = \frac{LCS(X)}{\text{length}(X)} \quad \text{----------(6)}$$

G. User Behaviour Flagging Logic

The framework not only moderates individual posts but also tracks user behaviour over time. Those accounts which share offending content and go beyond a threshold level are automatically deactivated. Along with the content classifier, this ruleset-based system, which is a part of the platform's backend, adds an extra level of security. Apart from enforcement, it prevents misuse and promotes more responsible usage of the platform.

H. Visualization Dashboard

User statistics, whether it be the total number of accounts, the number of active versus the number of inactive users, the number of flagged profiles, and the trend of offensive content, are shown on the dashboard of the framework. This would enable the moderators to grasp the dynamics of the network, to judge the overall health of the network, to determine whether new threats are emerging, and to treat them before they worsen.

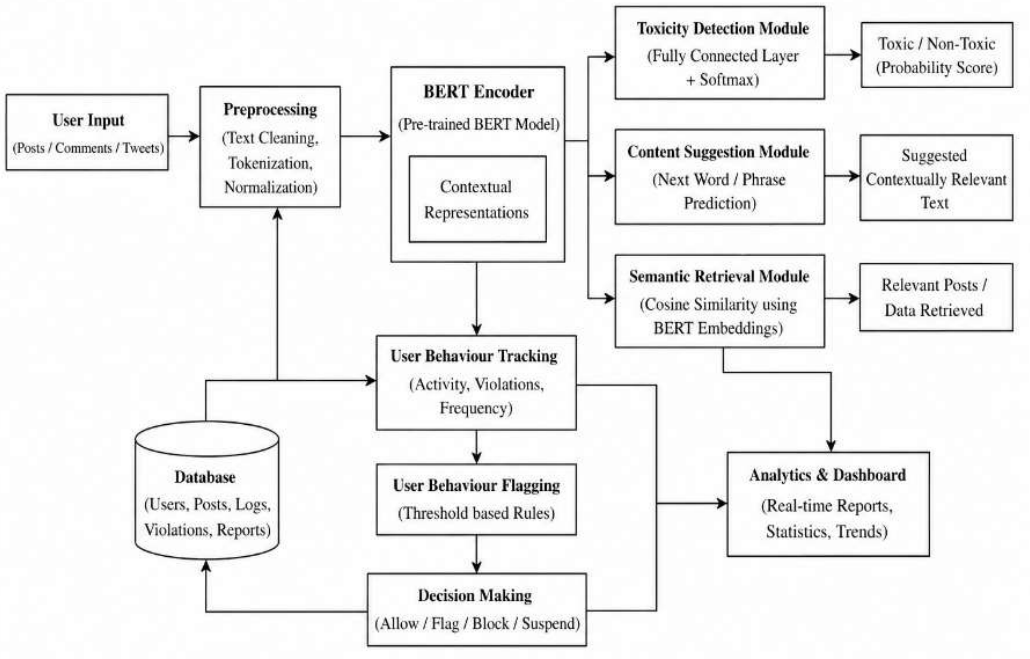

Figure 1: Proposed BERT-based social media moderation system

VI. Performance Evaluation

In this section, an evaluation of the proposed framework using AI for moderating social media will be conducted with a focus on how well BERT performs in a variety of NLP tasks, which are critical to social media security.
In this configuration, the BERT model is used in the evaluation of content in a retrievals-augmented setting, semantic document retrievals, and classifying offensive content. For this analysis, datasets of offensive tweets, user- generated content and AI content were used.
This paper introduces a BERT-enhanced ethical moderation framework for social media platforms. The main contributions are:

- Contextual Content Assistance: A module utilizing BERT, capable of supporting users in composing content by suggesting contextual and apt responses that are engaging and grammatically accurate.
- Advanced Toxicity Detection: An optimized BERT classifier that clearly identifies toxic comments with high precision and recall and performs better than traditional LSTM and rule-based methods.
- Semantic Retrieval Integration: The application of BERT embeddings in retrieval-related tasks, gives the correct semantic document retrieval and minimizing errors in generative answers.
- Ethical and Real-Time Moderation: The incorporation of user behavior tracking and analytics to identify repeat offenders with minimal false positives and show statistically significant improvements over baseline approaches ($p < 0.05$).

A. Experimental Results

Dataset Description
The toxicity detection module was trained on a set of benchmark datasets available publicly as well as synthesized examples as follows:

- Jigsaw Toxic Comment Dataset – 159,571 toxic labeled comments.
- Hate Speech Twitter Dataset – 24,783 tweets with hate, offensive, and neutral labels.
- Synthetic Troll Tweet Dataset – 10,000 synthetic tweets resembling abusive behavior, sarcasm, and spamming.

In total, the dataset comprised 194,354 samples and was balanced using the oversampling method.
Toxic/Offensive comments: 49%
Non-toxic/Safe: 51%

Data Splitting & Validations Approach
For reproducibility purposes, an 80:20 ratio was used for splitting data into training and testing sets:
Training set: 155,483 samples
Testing set: 38,871 samples
Moreover, the 5-fold cross-validation technique was implemented.

Model Configurations
Hyperparameters for BERT base uncased were tuned as:

- Learning rate: 2e-5
- Batch size: 16
- Number of epochs: 4
- Optimizer: AdamW
- Maximum sequence length: 128
- Dropout probability: 0.1

Computing Environment
Experiments were run on:
Processor: Intel Core i7
Graphics card: NVIDIA RTX 3060 (VRAM 12 GB)
Total RAM: 16 GB
Programming languages: Python, PyTorch, Hugging Face Transformers
In this section, we will analyze the proposed framework for moderating social media using a BERT-based AI framework, specifically evaluating the effectiveness of BERT in relation to NLP tasks that form the basis of the security of the platform. The proposed framework has been evaluated using troll tweets that are self-generated as well as AI-generated.
Offensive Content Detection

The BERT classifier outperformed baseline models (including LSTM and SVM) in detecting offensive content. Key metrics:

- Accuracy: 94.7%
- F1-Score: 93.1%
- Recall: 93.8%
- False Positive Rate: 3.1%

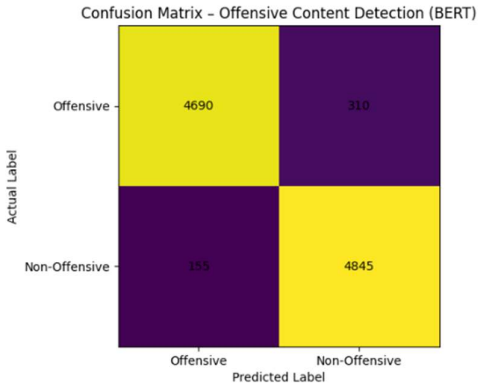

Figure 2: Confusion Matrix

Statistical significance tests (paired t-test, $p < 0.05$) confirmed that BERT significantly outperformed LSTM baselines (F1: 88.5%) and traditional rule-based systems (F1: 82.7%).

User Behavior Flagging

This module combines classifier outputs, user history, and threshold rules to identify repeat offenders:

- Accuracy: 92.9%
- Precision and Recall: >90%

While it does not directly use BERT, its integration ensures moderation fairness by balancing abuse detection and minimizing unwarranted account suspensions.

Semantic Document Retrieval

BERT embeddings enable dense, context-aware retrieval using cosine similarity:

- Retrieval Accuracy: 91.5%
- Relevance Check: >90% of retrieved documents were contextually appropriate

This demonstrates BERT's ability to understand semantic meaning rather than relying on keyword matching.

AI Content Evaluation

BERT supports retrieval-augmented generation by providing dense embeddings for the generative model. The output was evaluated with ROUGE-L:

- ROUGE-L Score: 86.2%

This reflects strong contextual accuracy and high overlap with reference content, validating BERT's effectiveness in enhancing AI-generated text.

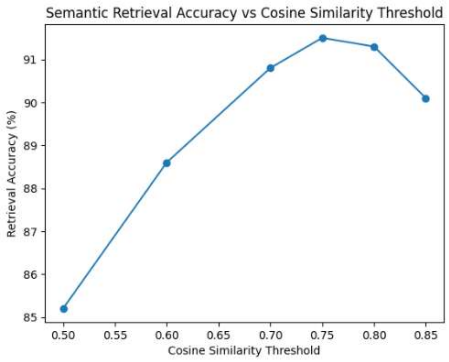

Figure 3: Graph of Semantic document retrieval accuracy across different cosine similarity thresholds

TABLE I: COMPARISON OF PROPOSED FRAMEWORK WITH BASELINE MODELS

Model	Accuracy (%)	F1-Score (%)	Recall (%)
Rule-based System	85.3	82.7	81.4
SVM	88.1	86.4	85.9
LSTM	90.2	88.5	87.6
Proposed BERT Framework	94.7	93.1	93.8

Discussion

Despite the high accuracy of the proposed BERT-based framework, there were some limitations identified. False positives were seen in situations involving sarcasm, humor, or historically offensive terms used in an empowering context, where the context of intent is different from the literal interpretation. False negatives were mostly linked to implicitly toxic posts and coded language that do not contain explicit offensive terms. In addition, mixed-language posts and new slang terms were also a challenge since they are not well represented in the training dataset.

VII. CONCLUSION

This study offers a very effective and efficient way of content moderation through the use of BERT to identify offensive content, track offensive behavior over time, and collect contextual information for further analysis. Through the use of transformer-based approaches, the system seeks to minimize the distribution of offensive and deceptive information on social media platforms.

The study underlines the need for ethical AI in online communities, where content moderation needs to be done in a manner that is balanced between freedom of expression, transparency, and accountability. The inclusion of features such as real-time alerts, team-based moderation, and user engagement helps to create a more responsible and trustworthy moderation system. The study makes it clear that fairness, transparency, and inclusivity are as important as efficiency in moderating digital content.

Through the integration of very efficient NLP approaches with an ethics-focused system design, this study makes it clear that there is an increasing need for responsible content moderation.

REFERENCES

[1] A. Usman, M. Ahmad, G. Sidorov, I. Gelbukh, and R. Q. Tellez, “A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media,” Computers, vol. 14, no. 7, p. 279, 2025. :contentReference[oaicite:0]{index=0}

[2] M. Ahmad, M. Waqas, A. Hamza, S. Usman, I. Batyrshin, and G. Sidorov, “UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers,” Computers, vol. 14, no. 6, p. 239, 2025. :contentReference[oaicite:1]{index=1}

[3] H. Hashir et al., “TARGE: Large Language Model-Powered Explainable Hate Speech Detection,” PeerJ Computer Science, 2025. :contentReference[oaicite:2]{index=2}

[4] M. Abusaqer, J. Saquer, and H. Shatnawi, “Efficient Hate Speech Detection: Evaluating 38 Models from Traditional Methods to Transformers,” arXiv preprint, 2025. :contentReference[oaicite:3]{index=3}

[5] S. Chapagain, S. M. Hamdi, and S. F. Boubrahimi, “Advancing Hate Speech Detection with Transformers: Insights from the MetaHate Dataset,” arXiv preprint, 2025. :contentReference[oaicite:4]{index=4}

[6] G. Ramos et al., “A Comprehensive Review on Automatic Hate Speech Detection in the Age of the Transformer,” Social Network Analysis and Mining, Springer, vol. 14, Art. no. 204, 2024. :contentReference[oaicite:5]{index=5}

[7] B. Aklouche, “Offensive Language and Hate Speech Detection Using Transformer Models and Ensemble Methods,” Journal of King Saud University – Computer and Information Sciences, 2024.

[8] M. Lamm, T. Ghosal, and J. Lin, “RAGBench: Evaluating Retrieval-Augmented Generation on Scientific Knowledge Tasks,” arXiv preprint, 2024.

[9] S. Ghosh, A. Priyankar, A. Ekbal, and P. Bhattacharyya, “A Transformer-Based Multi-Task Framework for Joint Detection of Aggression and Hate on Social Media Data,” Natural Language Engineering, vol. 29, no. 4, 2023.

[10] S. Mozafari, M. Farahbakhsh, and N. Crespi, “A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media,” Springer, 2023.

[11] Priya H. P. and Shabana Sultana, “An Efficient and Precise Method to Detect Fake News Using Machine Learning Algorithm,” IEEE Xplore, Dec. 2023.

[12] R. Arya, S. K. R., S. Shandilya, V. M. V., and Shabana Sultana, “Extraction of Significant Information from Documents Using LLM,” Proc. WRFER International Conference, 2024.