BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2645-2664
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2645-2664

EVALUATING ROBERTA AND GPT-BASED MODELS FOR SDG MULTICLASS TEXT CLASSIFICATION ACROSS DIFFERENT DOCUMENT LENGTHS

Uswatun Hasanah[1*], Agus Mohamad Soleh[2], Cici Suhaeni[3], Anwar Fitrianto[4]

[1,2,3,4] *Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University*
Jln. Meranti, Bogor, 16680, Indonesia

*Corresponding author's e-mail: * **21uswatun@apps.ipb.ac.id***

Article Info

Article History:

Received: 18th October 2025
Revised: 1st December 2025
Accepted: 17th March 2026
Published: 8th April 2026

Keywords:

Fine-tuning;
GPT;
RoBERTa;
SDGs;
Text Classification.

ABSTRACT

Multiclass text classification remains a difficult task, primarily due to semantic ambiguity and differences in input length. This study evaluates RoBERTa and GPT-based models for multiclass text classification, focusing on how prompting strategies and document length affect accuracy and robustness. Experiments were conducted using the OSDG Community Dataset, which contains approximately 15,000 labeled samples. The dataset was partitioned into four subsets based on input length: short, medium, long, and all combined. Three GPT variants (zero-shot, few-shot, and fine-tuned) were compared against a RoBERTa baseline. Fine-tuning was implemented via OpenAI's supervised API with prompt-response formatting. Performance was assessed through F1-score, precision, recall, and balanced accuracy. Fine-tuned GPT achieved the strongest results in all settings, with a macro F1-score of 0.9204 on the all-combined dataset, representing a 4.61% improvement over RoBERTa. Consistent gains were also observed across short (8.63%), medium (3.83%), and long (20.31%) texts. The largest improvement occurred on long documents, while medium-length inputs provided the most stable performance across models. These findings highlight the effectiveness of task-specific fine-tuning in enhancing GPT's capability to classify SDG-related texts across diverse input lengths.

How to cite this article:

U. Hasanah, A. M. Soleh, C. Suhaeni and A. Fitrianto, "EVALUATING ROBERTA AND GPT-BASED MODELS FOR SDG MULTICLASS TEXT CLASSIFICATION ACROSS DIFFERENT DOCUMENT LENGTHS", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2645-2664, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Founded by the United Nations in 2015, the UN Sustainable Development Goals (SDGs) serve as a comprehensive international framework of 17 interconnected aims to eradicate poverty, improve education, foster inclusive growth, and tackle climate change [1]. Nevertheless, efforts to achieve these goals have faced significant setbacks because of the global COVID-19 crisis, international political conflicts, and climate-related emergencies [2]. Currently, SDG tagging is manual, time-consuming, and subjective, which highlights the need for reliable automated approaches.

Automated SDG classification systems are increasingly in demand to support transparency, monitoring, and policy alignment [3]. Yet despite their value, these systems face technical challenges, including semantic ambiguity, thematic overlap, and variable document lengths. Recent findings indicate that document lengths can vary substantially across datasets, and such variation may influence model performance in multiclass settings. Han et al. [4] explicitly note that document lengths vary across datasets, and model performance can vary across length-varied corpora, highlighting that input-length heterogeneity is a non-trivial factor that affects classification robustness. Traditional approaches such as rule-based systems and keyword matching are insufficient to capture the contextual and thematic complexity of SDG-related texts [5], [6]. These challenges encourage the adoption of advanced models that better capture linguistic nuance and semantic structure. Among them, RoBERTa, an encoder-only transformer, has demonstrated strong performance in supervised classification tasks, especially when fine-tuned with domain-specific datasets [7]. Compared to BERT, RoBERTa benefits from longer training and dynamic masking while discarding the next-sentence prediction task, thereby enhancing its ability to capture contextual meaning [8]. However, RoBERTa still inherits the standard tokenization limit, and handling long documents often requires truncation, which may remove important contextual information and hinder model performance [4]. This suggests that RoBERTa's performance may be sensitive to document-length variation, especially when texts contain essential cues beyond its maximum sequence capacity.

The strong performance of RoBERTa in multiclass text classification has been widely documented. Sy et al. [8] reported consistently high accuracy across educational datasets, and Angin et al.[3] showed notable improvements on the OSDG Community Dataset compared to rule-based approaches. While RoBERTa requires labeled data, GPT models offer flexible inference via zero-shot and few-shot prompting, making them suitable for low-resource tasks. Recent studies highlight this versatility across domains. For example, Roumeliotis et al. [9] reported that GPT-4o outperformed BERT in hotel-review sentiment prediction, achieving 67% accuracy compared to 60.6% for BERT. In addition, unlike encoder-only models with strict input-length limits, GPT-based large language models can process substantially longer contexts. Sebők et al. [10] note that GPT-3.5 and GPT-4 support context windows ranging from 16k to 128k tokens, allowing them to process long inputs without truncation. Evidence from long-document classification further supports this capability: Trautmann [11] demonstrated that GPT-based prompting strategies, such as prompt chaining, can effectively classify legal documents of considerable length. Moreover, the effectiveness of GPT models extends to other specialized domains. For example, Nawab et al. [12] improved ICD code assignment in clinical notes by fine-tuning GPT-3.5 Turbo, demonstrating the viability of LLM-based classification in structured medical coding tasks.

Despite these advances, two gaps remain. First, there has been no systematic comparison of GPT-based model strategies, namely zero-shot, few-shot, and fine-tuned against RoBERTa within a multi-class, single-label SDG classification task. Second, the influence of input length on model robustness and accuracy has not been thoroughly evaluated. To address these gaps, this study evaluates four classification strategies: RoBERTa, GPT Zero-Shot, GPT Few-Shot, and Fine-Tuned GPT, applied to the OSDG Community Dataset with input-length stratification (short, medium, long, and all combined) and multiple performance metrics. This study contributes in two ways: (i) practical: providing evidence-based guidance for policymakers and institutions to adopt reliable and scalable SDG classification tools that enhance transparency and monitoring, and (ii) theoretical: advancing the understanding of how transformer-based models and large language models perform under different inference strategies and input-length conditions in a multi-class classification setting.

2. RESEARCH METHODS

This study follows a structured methodological pipeline covering dataset preparation, preprocessing, model implementation, and evaluation. The workflow is designed to assess the performance of encoder-based and generative transformer models under varying input-length conditions.

2.1 Data Collection

This study employs the OSDG Community Dataset (version 04.2024), curated by the OSDG team, to support Natural Language Processing (NLP) research related to the Sustainable Development Goals [13]. The dataset contains text segments sourced from public documents, including policy reports, governmental publications, and research abstracts. Texts were annotated by a diverse group of over 1,000 individuals spanning more than 100 countries, coordinated through the OSDG Community Platform, where each text sample was reviewed by at least 3 annotators, with a maximum of 9. Annotators provided binary decisions (accept/reject) for suggested labels, and an agreement score was calculated [13] as in Eq. (1):

$$Agreement = \frac{label_{accept} - label_{reject}}{label_{accept} + label_{reject}}. \quad (1)$$

Samples with low agreement (more negative than positive votes) were excluded to ensure label quality. The dataset comprises 15366 text segments, covering 16 out of 17 SDGs, as SDG 17 was not sufficiently represented in the corpus. For analysis of input granularity, the dataset was further stratified into short (<40 tokens), medium (40–99 tokens), and long (≥100 tokens) categories, in addition to the full dataset (All combined). This enables systematic evaluation of model robustness across varying input lengths.

2.1.1 Dataset Overview

This subsection presents an overview of the dataset and provides representative text samples used in the classification task. Example segments and their corresponding SDG labels are shown in Table 1 to illustrate the types and structures of the inputs used in this single-label multiclass setup.

Table 1. Example of Dataset Used in SDG Classification

No	Text	SDG Label
1	Another option would be to allow private practitioners to use the public clinic's facilities at low (or no) cost, from where they can bill the NHI for the patients that they see. This would still offer Korea's substantial number of solo practitioners the opportunity to reduce overhead and administrative costs for maintaining an independent practice. Unlike these countries, Korea lacks a large number of general practitioners trained, a strong professional identity among general practitioners, and high pay rates for general practitioners relative to specialists.	3
3	…	…
15366	This could be more effective than the preferential feed-in tariffs for infant industries. To boost wind energy development, the region could stimulate reindustrialization by leveraging its component manufacturing base, as Chicago has done. While surface geothermal energy is already well-established in the Paris-IDF region, there is considerable underexploited potential in deep geothermal energy, but it will require major investments.	9

2.1.2 Descriptive Statistics

To gain an overview of the dataset characteristics, we conducted an exploratory analysis focusing on the text length categories and the distribution of Sustainable Development Goal (SDG) labels. Fig.1 shows that medium-length texts (12,568) dominate the dataset of 15,366 samples, compared to short (1,959) and long (839) texts. Despite this imbalance, the stratification is sufficient for evaluating model performance across different input lengths.

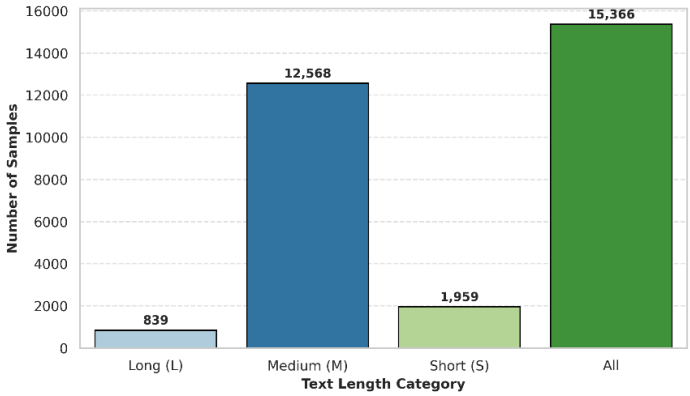


Figure 1. Number of Samples in Each Text Length Category

Fig. 2 illustrates the distribution of token lengths across all 15,366 text segments in the dataset. The distribution is right-skewed, with most samples ranging from 40 to 80 tokens and a long tail extending beyond 120 tokens. This confirms substantial variation in input lengths, reinforcing the need to evaluate model robustness under different sequence-length conditions.

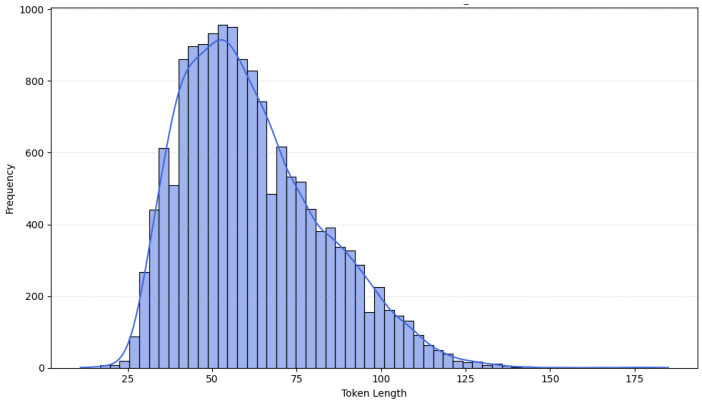


Figure 2. Distribution of Token Lengths in the OSDG Community Dataset

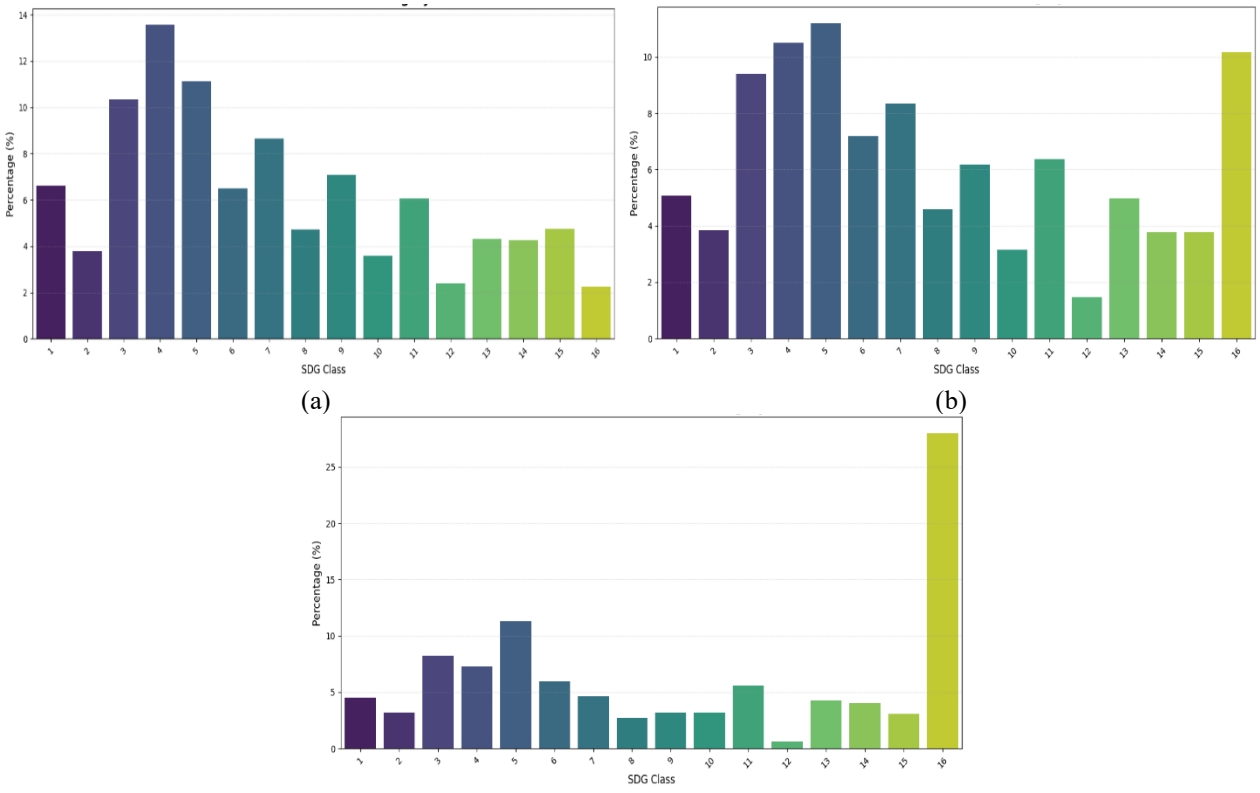


Figure 3. Distribution of SDG Classes by Text Length: (a) Short, (b) Medium, (c) Long

Fig.3 shows that the SDG class distribution differs across text length categories. Short texts contain higher proportions of SDG 4, 5, and 11; medium texts remain relatively balanced; and long texts are dominated by SDG 16. Several classes (e.g., SDG 2, 10, 12) appear consistently low across all groups. This confirms that the dataset is imbalanced both by class and by text-length category, justifying the need to evaluate model robustness across different input lengths.

2.2 Text Preprocessing

Following the data collection phase, a systematic text preprocessing workflow was implemented to prepare the dataset for modeling. In this study, RoBERTa and GPT-based models, grounded in tokenization and contextual representation learning, were employed; therefore, preprocessing was designed to preserve sentence structure and semantic integrity.

The preprocessing process included the following stages:

1. Cleansing: removing URLs, HTML tags, emojis, usernames, hashtags, and special characters.
2. Case Folding: all text was lowercased to ensure consistency and simplify token alignment.
3. Normalization: informal terms, abbreviations, and typographical inconsistencies were standardized using a curated lexicon, particularly to address domain-specific linguistic noise.

Unlike traditional pipelines, stopwords removal and stemming were omitted since prior studies show they can reduce transformer performance by disrupting contextual coherence [14]. Moreover, recent large-scale surveys emphasize that although transformers appear robust, their sensitivity to preprocessing choices remains high, reinforcing the need for careful, and minimal intervention during input preparation [15].

The dataset was thus preserved in a cleaned, lowercased, and semantically intact form compatible with both RoBERTa's and GPT's tokenizers. The structural flow of the preprocessing procedure is encapsulated in Fig. 4, and Table 2 provides example text segments before and after preprocessing.

Figure 4. Text Preprocessing for Transformer-based Models (RoBERTa and GPT-based Modes)

Table 2. Sample of Preprocessed Text Segments for SDG Classification

No	Before Pre-processing	After Preprocessing
1	Another option would be to allow private practitioners to use the public clinic's facilities at low (or no) cost from where they can bill the NHI for the patients that they see. This would still offer Korea's substantial number of solo practitioners the opportunity to reduce overhead and administrative costs for of maintaining an independent practice. Unlike these countries, Korea does not have a large number of doctors trained as general practitioners, a strong professional identity for general practitioners, and high pay rates for general practitioners relative to specialists.	Another option would be to allow private practitioners to use the public clinics facilities at low or no cost from where they can bill the nhi for the patients that they see this would still offer koreas substantial number of solo practitioners the opportunity to reduce overhead and administrative costs for maintaining an independent practice unlike these countries korea does not have a large number of doctors trained as general practitioners a strong professional identity for general practitioners and high pay rates for general practitioners relative to specialists
2	….	….
15366	This could be more effective than the preferential feed-in tariffs for infant industries. To boost the development of wind energy, the region could stimulate its reindustrialisation by taking advantage of its component manufacturing base, as Chicago has done. While surface geothermal energy is already well-established in the Paris-IDF region, there is considerable underexploited potential in deep geothermal energy, but it will require major investments.	This could be more effective than the preferential feed in tariffs for infant industries to boost the development of wind energy the region could stimulate its reindustrialisation by taking advantage of its component manufacturing base as chicago has done while surface geothermal energy is already well established in the paris idf region there is considerable underexploited potential in deep geothermal energy but it will require major investments

2.3 Dataset Splitting

A stratified train-test split was performed after preprocessing, with 70% of the samples assigned to training and 30% to testing, while preserving the SDG category balance. This helps ensure the training set is representative while enabling fair evaluation on unseen data. The dataset is partitioned as follows [16]:

1. 70% for training: Employed to fine-tune RoBERTa and train the fine-tuned GPT model, ensuring adequate exposure to examples from all SDG labels.
2. 30% for testing: Held out for evaluating all classification modes, including RoBERTa, GPT zero-shot, few-shot, and fine-tuned, covering all input-length categories (short, medium, long, and all combined), ensuring consistent evaluation across all architectures and prompting strategies.

By performing splitting separately within each length-defined subset, all models are trained and evaluated on data that reflect the characteristics of the specific input-length group. This ensures consistent and fair comparisons across architectures and supports a controlled analysis of model robustness under different document lengths.

2.4 Classification Models

This study compares two transformer-based classification strategies for mapping text segments to Sustainable Development Goals (SDGs):

1. Fine-Tuned RoBERTa Model

   This model leverages a RoBERTa-based architecture and is adapted through supervised learning on annotated data to optimize its contextual representations for SDG classification [3].

2. GPT-Based Model

   This includes three approaches: Zero-Shot, Few-Shot, and GPT Finetuning. Zero-shot and few-shot variants use prompt-based inference without retraining model weights [17], [18], while GPT Finetuning is adapted via supervised prompt-response examples [19], [20], [21].

2.5 Model Configuration

To evaluate performance on multi-class classification, we employed three strategies: RoBERTa, fine-tuned GPT, and GPT prompting (few-shot and zero-shot). This section details each model's configuration.

2.5.1 RoBERTa

RoBERTa model, proposed by [22], is an encoder-only transformer architecture built upon the original BERT framework [23]. The model comprises several stacked encoder layers that incorporate self-attention, feed-forward networks, residual connections, and normalization to effectively extract deep contextual information from text [22].

RoBERTa distinguishes itself from the BERT framework by eliminating the next-sentence prediction objective and incorporating dynamic masking, larger batch sizes, and longer pre-training, all of which improve downstream performance [24], [25]. The Roberta-base model and RobertaTokenizer from the Hugging Face Transformers library were fine-tuned in this study. Tokenization was performed with truncation and padding enabled, and maximum sequence length was capped at 256 tokens to ensure efficiency while accommodating all text samples [8]. Class labels were converted to zero-based integer indices to align with PyTorch.

A total of 10 training epochs were used during the fine-tuning stage, with the AdamW optimizer using a fixed learning rate of 1×10^{-5} [3]. The choice of a relatively small learning rate follows previous findings indicating that lower learning rates help prevent catastrophic forgetting and ensure stable convergence during the fine-tuning of large transformer-based models. The AdamW optimizer was chosen because it has been shown to effectively train transformer-based models [26]. By separating weight decay regularization from the gradient update, AdamW reduces overfitting and enhances the model's ability to generalize to unseen data [27]. A batch size of 32 was employed to balance computational efficiency and gradient stability, as larger batches, although potentially reducing generalization, typically lead to faster convergence and more efficient training [28]. The training process was conducted for 10 epochs [29], a choice empirically justified by its capacity to deliver adequate parameter updates conducive to convergence while mitigating the risk of

overfitting. Moreover, the sequence length constraint was set to 256 tokens, a compromise that preserved contextual depth without incurring excessive computational demand [22]. Lastly, the optimization objective incorporated a cross-entropy loss and a softmax activation function to effectively address the multiclass classification challenge [30].

2.5.2 GPT-Based Model

Known as Generative Pre-trained Transformer (GPT), this model constitutes a large-scale, decoder-centric transformer architecture introduced by [31] and subsequently expanded into more advanced versions, including GPT-3 and GPT-4 [32]. Built on the foundational transformer design proposed by [33], GPT uses a decoder-based architecture composed of hierarchically stacked transformer layers that integrate multi-head self-attention, position-wise feed-forward networks, residual pathways, and layer normalization. Collectively, these architectural components enable GPT to perform effective natural language processing and generation, yielding coherent, contextually appropriate outputs [34].

Within this structural framework, the language model operates as a multi-layer transformer decoder that applies multi-headed self-attention to input sequences, followed by feed-forward layers to compute probabilistic output distributions over target tokens [32]. The canonical configuration typically comprises 12 decoder layers utilizing masked self-attention, 768-dimensional hidden representations, and 12 attention heads, while the feed-forward layers employ 3072-dimensional intermediate states, and optimization is achieved via the Adam algorithm [31].

2.5.2.1. GPT Fine-Tuning

GPT-4.1- mini is fine-tuned to improve its performance on domain-specific classification tasks, particularly in specialized application areas. This process enables a general-purpose pre-trained model to specialize in specific tasks by adapting its parameters based on labeled training examples, thereby aligning its outputs with the requirements of SDG classification [35], [36]. The training corpus was derived from the OSDG Community Dataset, comprising 15,366 labeled samples, which were partitioned into 70% for training and 30% for testing. The fine-tuning workflow consisted of two stages: (i) prompt-completion formatting and (ii) model fine-tuning via the OpenAI API.

Stage 1: Prompt-Completion Formatting

To facilitate fine-tuning, all samples were converted into prompt–completion pairs according to the guidelines in the OpenAI documentation [37]. Each pair consisted of an instruction-like prompt, user input, and the expected SDG label as the completion. A fixed instruction template was employed to guarantee consistency across the dataset:

```
System: You are an intelligent classification system that maps texts to any of
     the 16 UN Sustainable Development Goals (SDGs), numbered 1 through 16. Given
     the following text, determine which SDG (1 to 16) best represents its main
     theme. Respond with only a single integer between 1 and 16. Do not include
     any explanation or extra text, just the number.
User: Text: "[TEXT]"
Assistant: [SDG_LABEL]
```

The data were serialized into JSONL format, with each line representing a single training example ready for fine-tuning.

Stage 2: Fine-Tuning Process

After structuring the dataset, the JSONL file was uploaded to the OpenAI File API with a unique identifier. Fine-tuning on the GPT-4.1-mini model followed a supervised learning approach, updating model parameters from labeled prompt–completion pairs to improve task-specific performance.

The training configuration used two epochs, following the setup from previous fine-tuning experiments [38], with a batch size of 32 as reported in similar work [39]. A comparatively low epoch count was selected to reduce computational cost and avoid excessive training iterations, which can lead to overfitting in transfer learning settings [40]. The batch size of 32 was chosen as a standard configuration commonly adopted in transformer fine-tuning, providing a practical balance between gradient stability and training efficiency. A learning rate multiplier of 0.1 was used. The value was selected empirically as the most stable during

preliminary trials, as prior work has empirically shown that such configurations can slightly improve validation performance [41]. Finally, evaluation was conducted using the Chat Completion API with a temperature of 0, a configuration that ensures deterministic outputs and consistent evaluation conditions [17]. A schematic representation of the procedural framework is delineated in Algorithm 1.

Algorithm 1. Fine-tuning GPT-4.1-mini for SDG Classification

Input : OSDG dataset in prompt-completion JSONL format (15,366 samples).

Output : Fine-tuned GPT-4.1-mini model $f_{\boldsymbol{\theta}}$.

1. Initialize the pre-trained model $f_{\boldsymbol{\theta}}$ (GPT-4.1-mini).
2. For each training pair (p_i, c_i):
   a. Encode the prompt p_i.
   b. Predict the output distribution $\hat{c}_i = f_{\boldsymbol{\theta}}(p_i)$.
   c. Compare $\hat{c}_i$ with the target label c_i.
3. Adjust model parameters iteratively using mini-batches of size 32.
4. Repeat the training process for two training epochs.
5. Store the resulting fine-tuned model identifier $f_{\boldsymbol{\theta}*}$ for inference.

2.5.2.2. GPT Prompting (Zero-shot and Few-shot Prompting)

The model's in-context reasoning ability was evaluated using GPT-4.1- mini via OpenAI's ChatGPT API. These approaches operate without parameter tuning, leveraging the model's pre-trained knowledge [17], [42]. In the zero-shot setup, the model receives an instructional prompt and input text, following OpenAI's template for classification tasks:

```
System: You are an intelligent classification system that maps texts to any of
     the 16 UN Sustainable Development Goals (SDGs), numbered 1 through 16. Given
     the following text, determine which SDG (1 to 16) best represents its main
     theme. Respond with only a single integer between 1 and 16. Do not include
     any explanation or extra text, just the number.
User: Text: "[TEXT]"
Assistant: [SDG_LABEL]
```

For the few-shot configuration, we embedded five labeled exemplars per SDG (totaling 80 in-context examples) directly into the prompt. The examples were drawn from the training set using a stratified, fixed-sampling approach to ensure representativeness and reproducibility [43]. The unlabeled test instance was appended at the end. The average prompt length was ~2,800 tokens while remaining fully compatible with the GPT-4 context window, thereby avoiding truncation. All GPT prompting was conducted with temperature = 0 to ensure deterministic outputs, following OpenAI's best practices [44]. Inference was performed on a held-out stratified test set (30% of OSDG data; N = 15,366), consistent with the split used in the fine-tuned model. The procedure is in Algorithm 2.

Algorithm 2. GPT Prompting for SDG Classification

Input:

Test set $\{(x_j, y_j)\}_{j=1}^{M}$ with $M = n_{test}$

x_j: input text instance (document, review, or report)

y_j: ground-truth SDG label, $y_j \in \{1, \ldots, 16\}$

Prompt template T

Exemplar set E (few-shot only)

Output: Predicted labels $\hat{y}_j$ for each test sample

1. For each test instance (x_j, y_j):
   a. If zero-shot: construct prompt $p_j = T(x_j)$.
   b. If few-shot: construct prompt $p_j = T(E \cup \{x_j\})$.
2. Submit p_j to GPT-4.1-mini using Chat Completions API with temperature = 0.
3. Parse the output as a single integer label $\hat{y}_j \in \{1, \ldots, 16\}$.
4. Collect predictions $\{\hat{y}_j\}_{j=1}^{M}$.
5. Compare $\hat{y}_j$ with ground truth y_j to compute macro-level metrics (F1-score, Balanced Accuracy, Precision, Recall).

2.6 Evaluation Metrics

The classification results were assessed using four metrics derived from the confusion matrix, where each entry M_{ij} indicates the count of instances from class i predicted as class j, as in Eqs. (2), (3), (4), and (5) [45], [46]:

$$Balanced\ Accuracy = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i}, \tag{2}$$

$$Precision_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FP_i}, \tag{3}$$

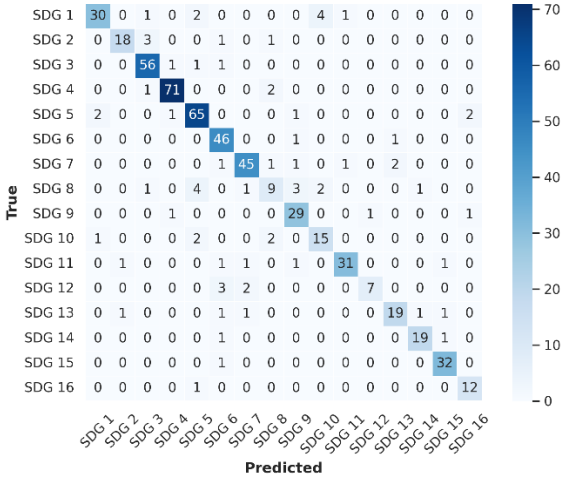
$$Recall_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{TP_i}{TP_i + FN_i}, \tag{4}$$

$$F1 - Score_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \times TP_i}{2 \times TP_i + FP_i + FN_i}. \tag{5}$$

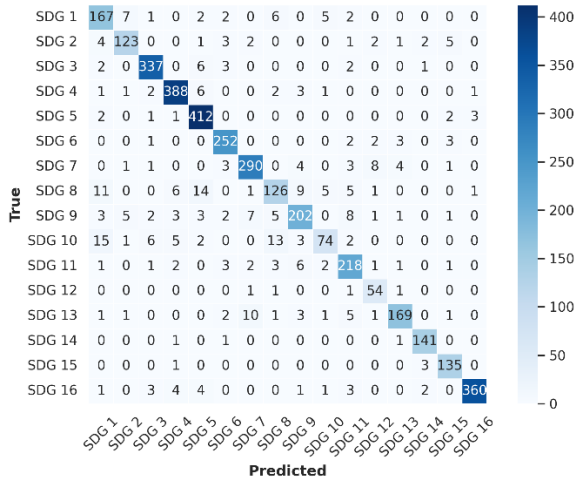
3. RESULTS AND DISCUSSION

3.1 RoBERTa Implementation

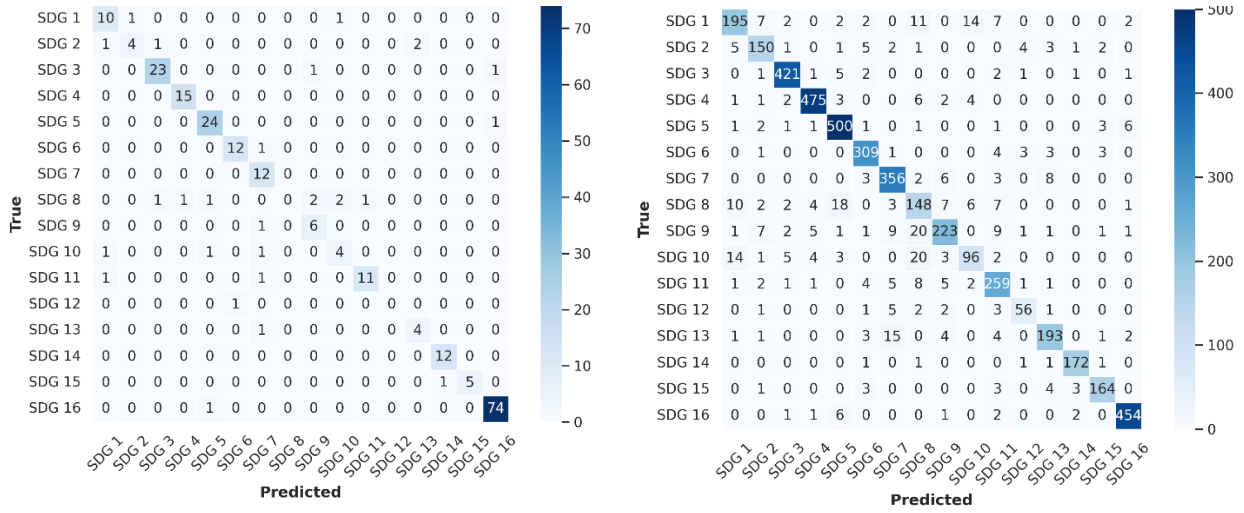
To examine RoBERTa's behavior across different input-length categories, we evaluated its performance separately on short, medium, long, and all-combined text groups using identical train–test splits. The evaluation uses four metrics, namely balanced accuracy, macro F1, macro precision, and macro recall.



(a)



(b)



(c) (d)

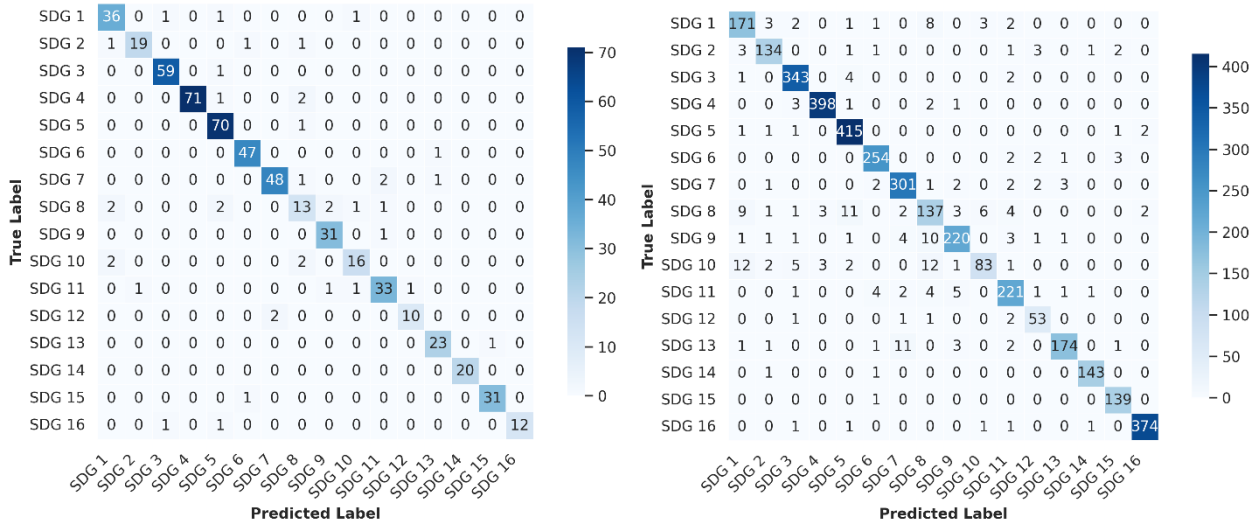
Figure 5. Confusion Matrices on the Test Set Across Input-Length Categories for RoBERTa: (a) Short, (b) Medium, (c) Long, (d) All combined.

Fig. 5 shows RoBERTa's confusion matrices for short, medium, long, and all combined inputs. Medium texts produce the strongest diagonals, with high true-positive counts for major SDGs such as 3, 4, 5, 6, 7, and 15, indicating stable separation across classes. Short inputs remain fairly accurate but display greater confusion, especially for lower-resource goals like SDGs 8, 9, 10, and 12, reflecting the limited contextual cues in brief texts.

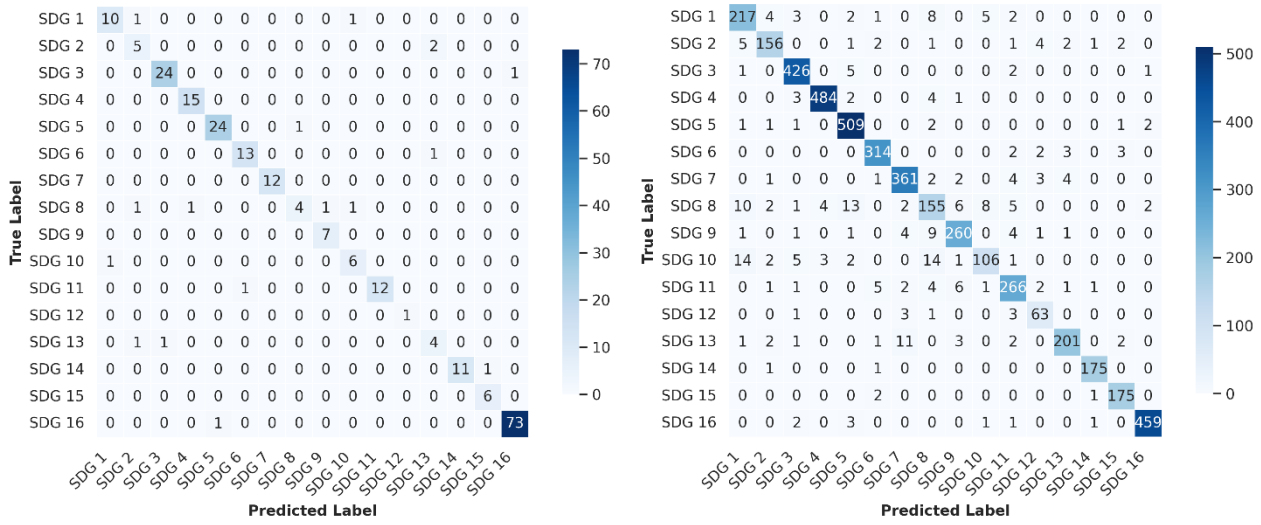
Long inputs yield the weakest patterns because only a small number of documents fall into this category, resulting in sparse diagonals and more split predictions for SDGs with very few samples. The combined matrix mirrors the dataset imbalance: high-frequency SDGs are predicted reliably, whereas low-resource SDGs, such as 10, show greater dispersion. Overall, RoBERTa performs most consistently on medium-length texts, with accuracy decreasing for short and especially long inputs.

3.2 Fine-Tuned GPT

The fine-tuned GPT model was evaluated using the same train–test split and input-length categories as RoBERTa. The assessment focuses on the model's prediction outputs, including class-level performance and error distribution.



(a) (b)



(c) (d)

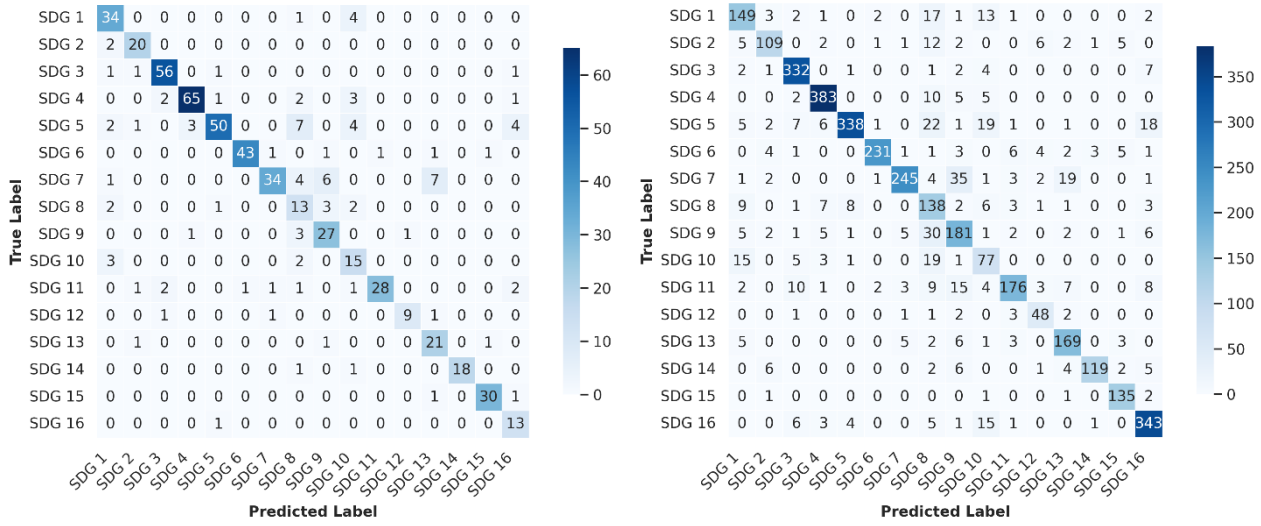
Figure 6. Confusion Matrices on the Test Set Across Input-Length Categories for Fine-Tuned GPT: (a) Short, (b) Medium, (c) Long, (d) All combined.

Fig. 6 shows across all input-length subsets, the model demonstrates strong predictive accuracy for high-frequency SDGs such as SDG 3, 4, 5, and 16, as indicated by the clear diagonal patterns in the confusion matrix. Medium-length texts perform best, achieving high accuracy with minimal misclassification, particularly for SDG 5 and SDG 16. This suggests that medium-length inputs provide contextual richness. In comparison, short texts exhibit more frequent misclassification, especially within semantically overlapping categories such as SDG 1, 2, and 8. Although SDGs 3 and 4 perform well, short texts show wider errors, reflecting difficulty distinguishing nuanced categories with limited context. For example, SDG 2 is often conflated with SDG 1, and SDG 8 tends to overlap with SDG 1 or SDG 13.

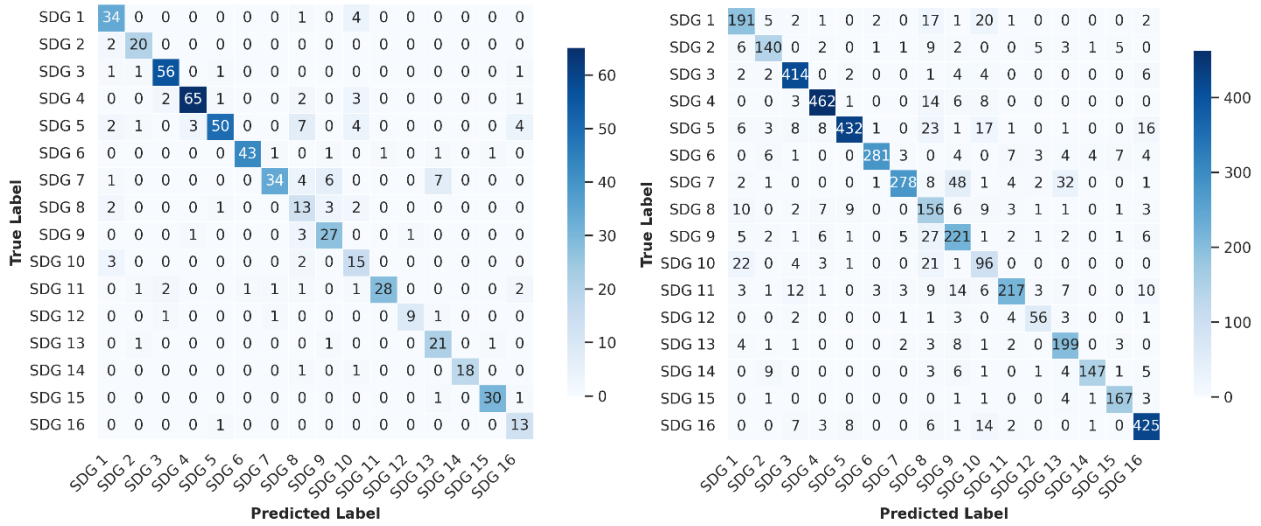
For long inputs, the model remains accurate on dominant goals such as SDG 16, but overall predictions are sparser. This may result from truncation, saturation, or a diluted signal-to-noise ratio, where important features are buried in a long context. In the combined evaluation across all input lengths, the model continues to perform well on frequently occurring categories, including SDG 3, SDG 5, and SDG 16. However, misclassifications persist for less frequent goals, including SDG 8, 9, 10, 12, and 13. These labels are often confused with similar goals (e.g., SDG 1 or 8), highlighting challenges posed by class imbalance and semantic overlap.

3.3 Zero-Shot GPT Performance on SDG Classification

GPT, in its zero-shot configuration without task-specific adaptation, exhibits different performance patterns than its fine-tuned counterpart.



(a) (b)



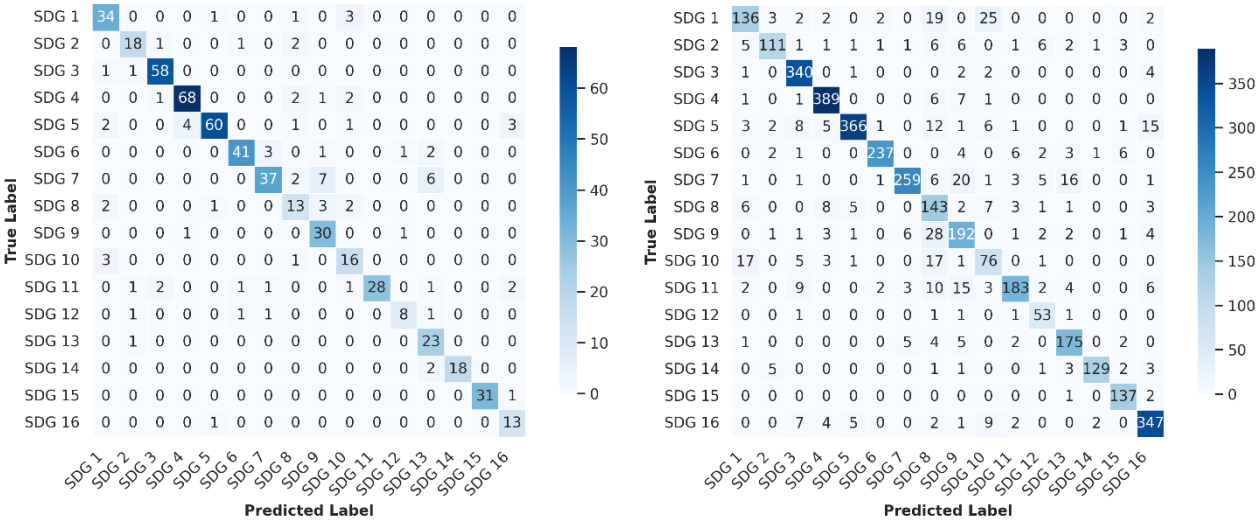
(c) (d)

Figure 7. Confusion Matrices on the Test Set Across Input-Length Categories for Zero-Shot Prompting GPT: (a) Short, (b) Medium, (c) Long, (d) All combined.

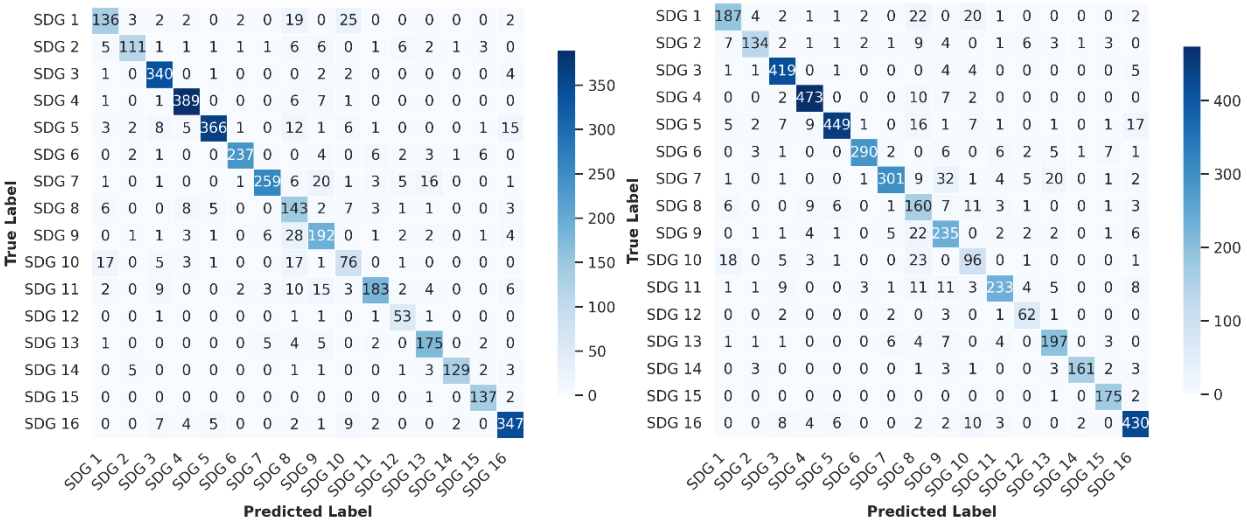
As shown in Fig. 7, it performs reasonably well on frequent, semantically distinct goals such as SDGs 3, 4, 5, and 16. However, it struggles with goals that have overlapping themes, such as SDG 1, 2, and 8, where many instances of SDG 1 are misclassified as SDG 8 or 10. These errors indicate that the model depends on surface-level cues rather than deep semantic understanding, limiting its ability to distinguish closely related goals. Although diagonal structures remain evident, the overall prediction density is lower than that of the fine-tuned GPT, reflecting diminished precision and greater uncertainty. SDG 7, for example, is often misclassified as SDG 13 or 10 due to shared environmental or infrastructure-related keywords. These findings highlight the limitations of zero-shot GPT in nuanced SDG classification and emphasize the need for fine-tuning or few-shot approaches, especially when dealing with low-resource labels.

3.4 Few-Shot GPT Performance across Input Lengths

Compared to zero-shot, few-shot prompting significantly enhances model calibration and label distinction, as seen in Fig. 10, where confusion matrices show clearer and denser diagonal patterns, reflecting improved prediction accuracy.



(a) (b)



(c) (d)

Figure 8. Confusion Matrices on the Test Set Across Input-Length Categories for Few-Shot Prompting GPT: (a) Short, (b) Medium, (c) Long, (d) All combined.

Fig. 8 shows that few-shot prompting performs best on medium-length inputs, yielding accurate predictions for frequent goals such as SDG 3, 4, 5, and 16, with reduced confusion across overlapping categories. Gains are also noted for SDG 6, 7, and 10, showing that representative examples help clarify context. However, short texts remain challenging; despite slight improvements, misclassifications persist for semantically similar goals, such as SDG 1, 2, 8, and 13, due to limited context.

Long inputs show no significant advantage over medium ones. The confusion patterns remain similar, suggesting performance saturation likely caused by diluted relevant information or token limit constraints. Across all input lengths, accuracy is high for dominant goals (SDG 5, 16) but struggles with rare or semantically close labels. Frequent mix-ups, such as SDG 8 with SDG 9 or SDG 1 with SDG 2, highlight the persistent challenges of semantic overlap and class imbalance, even with few-shot prompting.

3.5 Representative Text-Level Prediction Cases

To complement the quantitative evaluation and provide concrete evidence for the classification patterns observed in the confusion matrices, Table 3 presents selected examples that reveal distinct success and failure patterns across models.

Table 3. Sample of Preprocessed Text Segments for SDG Classification

No	Text	True SDG	Zero-Shot	Few-Shot	Fine-Tuned	RoBERTa
1	many technology however still experimental stage discusses new technology could potentially shorten last mile tor household smes remote area suggests investment could help close existing gap partnership effective within well designed regulatory environment however resource currently absent short supply many smes especially remote location	9	9	9	9	7
2	since road death steadily decreased fluctuation halved korea counted road death decrease compared level population million million registered vehicle around vehicle per inhabitant korea road network total includes motorway	11	3	3	11	11
3	collection medical waste transported treatment facility sterilized hydroclave transported disposal site disposed together general municipal waste considering infectious waste represents per cent total medical waste manage quarter infectious waste generated albania seven hospital house sterilization unit dispose waste independently	12	3	3	3	12

The examples in Table 3 highlight how the four models handle different thematic cues in SDG-related texts. In the first example, all GPT variants correctly predicted SDG 9, while RoBERTa misclassified the text into SDG 7. This suggests that GPT models captured the broader context of innovation and industrial development, whereas RoBERTa was more affected by energy-related keywords that commonly co-occur with SDG 7. In the second example, RoBERTa and Fine-tuned GPT correctly identified SDG 11 due to clear references to transportation safety and urban infrastructure, whereas Zero-Shot and Few-Shot GPT misclassified the text as SDG 3, likely because health-related terms dominated their decisions. The third example shows the opposite pattern: RoBERTa correctly predicted SDG 12 by recognizing waste-management cues, while all GPT variants predicted SDG 3, reflecting their sensitivity to recurring medical and healthcare terminology within the text. These cases demonstrate that GPT models tend to prioritize dominant thematic signals, whereas RoBERTa is more influenced by specific keyword patterns, resulting in complementary strengths across different SDG contexts.

3.6 Model Performance and Stability Evaluation

This study evaluated the performance of four classification approaches applied to SDG-related texts. Fig. 9 presents a bar chart comparing their F1-scores across four input-length subsets.

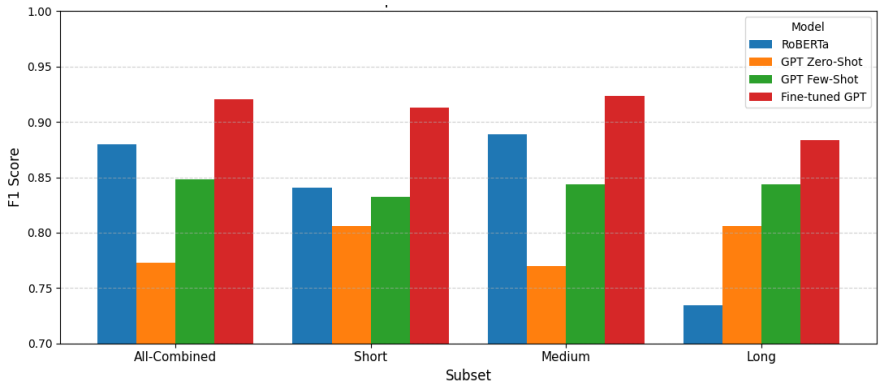


Figure 9. Bar Chart Illustrating the Macro F1-Score Comparison Among Roberta, Gpt Zero-Shot, Gpt Few-Shot, and Fine-Tuned GPT Across Four Input-Length Subsets

Fig. 9 compares macro F1-scores across input-length subsets and shows that Fine-tuned GPT achieves the strongest class-balanced performance across all groups. RoBERTa ranks second-best, with relatively stable macro F1 scores, particularly in the all-combined and medium subsets, indicating that it can handle both major and minor SDG classes more reliably than zero-shot or few-shot prompting. Few-Shot GPT consistently performs better than Zero-Shot across all subsets, showing that even a small number of examples helps the model distinguish between SDGs with greater class balance. Zero-Shot GPT produces the lowest macro F1 values, with the shortest bars across all four groups, reflecting weaker performance on underrepresented SDGs when no examples are provided.

Fine-tuned GPT achieved the strongest performance across all input-length subsets, as summarized in Table 4. On the all-combined dataset, it produced the highest macro F1-score (0.9204) along with strong precision (0.9237), recall (0.9185), and balanced accuracy (0.9185). This advantage remained consistent across short (F1 = 0.9130), medium (F1 = 0.9232), and long texts (F1 = 0.8839), indicating that the model generalizes well across varying document length and contextual richness.

Table 4. Evaluation Metrics for Four Classification Approaches on SDG Text Data Across all length.

Dataset	Evaluation	Methods			
		RoBERTa	GPT-Zero Shot	GPT-Few Shot	Fine-tuned GPT
SDGs All	Balance Accuracy	0.8759	0.8251	0.8525	0.9185
	F1 Macro	0.8798	0.7728	0.848	0.9204
	Precision Macro	0.8854	0.7768	0.8496	0.9237
	Recall Macro	0.8759	0.7765	0.8525	0.9185
SDGs Short	Balance Accuracy	0.8347	0.8254	0.8451	0.9072
	F1 Macro	0.8404	0.8059	0.8326	0.913
	Precision Macro	0.8548	0.8085	0.8341	0.9206
	Recall Macro	0.8347	0.8254	0.8451	0.9072

Dataset	Evaluation	Methods			
		RoBERTa	GPT-Zero Shot	GPT-Few Shot	Fine-tuned GPT
SDGs Medium	Balance Accuracy	0.8906	0.8225	0.8493	0.9215
	F1 Macro	0.8891	0.7698	0.8436	0.9232
	Precision Macro	0.8919	0.7739	0.8454	0.9271
	Recall Macro	0.8906	0.7741	0.8493	0.9215
SDGs Large	Balance Accuracy	0.7519	0.8254	0.8493	0.8904
	F1 Macro	0.7347	0.8059	0.8436	0.8839
	Precision Macro	0.7289	0.8085	0.8454	0.885
	Recall Macro	0.7519	0.8254	0.8493	0.8904

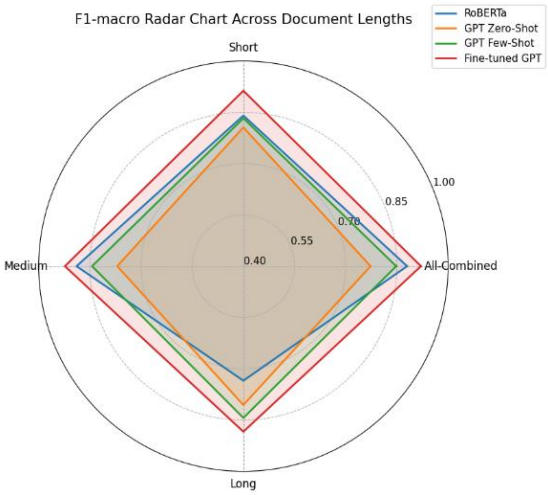


Figure 10. Radar Chart Visualization of f1-Macro Stability Across all Combined, Short, Medium, and Long Subsets for Each Classification Method.

Fig. 10 illustrates the macro F1-score comparison across the four input-length categories. Fine-tuned GPT takes the largest and most uniform shape, demonstrating consistently strong performance across text lengths. The model maintains high scores across all-combined, short, medium, and long subsets, indicating that fine-tuning provides strong generalization even when contextual details vary. RoBERTa ranks second but shows greater variation across text lengths. Its highest macro F1-score occurs in the medium subset, remains competitive in the all-combined subset, drops slightly for short inputs, and reaches the lowest value on long texts. GPT Few-Shot forms a smaller but relatively stable profile, outperforming the zero-shot setting across all subsets. GPT Zero-Shot consistently produces the smallest polygon, indicating limited performance without example-based guidance or fine-tuning.

Text classification remains a central component of NLP for sustainable development, yet previous studies have not thoroughly examined how variations in input granularity influence the performance of GPT-based models. This study fills that gap by systematically evaluating zero-shot, few-shot, and fine-tuned GPT models on SDG texts of varying lengths. Prior studies, including those by [47] and [48], demonstrate that fine-tuned BERT-based models often surpass prompting-based methods in binary or abstract tasks. Nonetheless, these works did not account for input length variation or benchmark against decoder-based architectures, such as GPT-based models.

The results confirm that Fine-tuned GPT remains the strongest performer across all input-length categories, achieving the highest macro F1-scores on every subset. Its performance is consistently superior across all input lengths, and although the score decreases on long texts, it still outperforms the other approaches. Zero-shot GPT produced the weakest results, particularly because it struggled to handle semantically overlapping SDGs without task-specific examples. Few-shot prompting improved performance over Zero-shot and produced a more balanced profile, but it generally remained below RoBERTa. RoBERTa performed reliably, especially on medium-length texts, yet still trailed Fine-tuned GPT across all input lengths.

This aligns with prior work, such as [20], which reported a 10.6% boost in performance when fine-tuning GPT-3.5 over BERT in the domain of automated educational scoring. Study [18] investigated input-length variation using the SDGi corpus but did not evaluate GPT-based model strategies. Similarly, studies

such as [9] explored tuned models without evaluating performance shifts across input granularities. The superiority of fine-tuned GPT can be explained by domain adaptation, prompt-response supervision, and richer contextual embeddings, which enable the model to capture nuanced SDG semantics beyond prompting. Meanwhile, RoBERTa's stability on medium texts highlights the strengths of encoder-only architectures for dense but not overly long inputs. This complementary pattern underscores the primary contribution of this research by providing a systematic evaluation of GPT strategies and RoBERTa across varying input-length conditions, addressing a gap in prior SDG classification studies and offering evidence that fine-tuned LLMs can serve as tools for sustainability-focused NLP applications.

3.7 Computational Cost vs Accuracy Trade-off

While Fine-tuned GPT achieves the highest accuracy, its deployment requires recurring API usage, which introduces operational costs for public institutions. The improvement over RoBERTa in the all-combined subset reaches 4.61% in macro F1, with gains ranging from 3.83% to 20.31% across specific input-length categories. Whether these improvements justify continuous API expenses depends on the institutional context. For high-stakes policy applications, where misclassification can lead to incorrect thematic alignment or downstream decision-making errors, even a 3–5% performance gain may offer substantial practical value. In contrast, for routine or large-volume classification tasks with lower sensitivity to marginal differences in accuracy, a locally hosted RoBERTa model may be a more cost-efficient option, particularly because it incurs no additional inference cost once deployed. These findings indicate that Fine-tuned GPT is most suitable for government institutions requiring high reliability across diverse document lengths, while RoBERTa remains a viable and cost-effective baseline for resource-constrained environments.

4. CONCLUSION

This study compared RoBERTa with three GPT-based approaches, namely zero-shot, few-shot, and fine-tuned GPT, for multiclass SDG text classification across short, medium, long, and all-combined inputs. Fine-tuned GPT consistently achieved the strongest performance, with macro F1-scores above 0.88 and improvements of 3.8-20.3% over RoBERTa across the different input-length categories. RoBERTa remained the second-best model and performed most reliably on medium-length texts, while zero-shot and few-shot prompting produced weaker results, particularly for semantically overlapping and low-resource SDGs. These findings highlight the value of task-specific fine-tuning for achieving accurate and stable SDG classification across variable document lengths, supporting the development of automated monitoring systems for policy and institutional use. Future research may explore multi-label or hierarchical SDG classification, multilingual SDG corpora, significance testing, and multi-fold validation to improve reliability assessment, and the integration of explainability techniques to enhance transparency and auditability in sustainability-oriented decision-making.

Author Contributions

Uswatun Hasanah: Conceptualization, Methodology, Data Curation, Software Implementation, Model Development, Formal Analysis, Investigation, Visualization, Writing – Original Draft, Writing – Review and Editing. Agus Mohamad Soleh: Supervision, Conceptual Guidance, and Manuscript Review. Cici Suhaeni: Supervision, Conceptual Guidance, and Manuscript Review. Anwar Fitrianto: Statistical Supervision, Methodological Validation, and Manuscript Review. All authors discussed the results and contributed to the final manuscript.

Funding Statement

This project received funding from the Directorate General of Research and Development, Ministry of Education, Science, and Technology, under the Implementation Contract of the 2025 Research Program, Contract Number: 006/C3/DT.05.00/PL/2025.

Acknowledgment

The authors would like to express their gratitude to IPB University for the academic support provided during the research process. The authors also extend sincere appreciation to the reviewers and editorial team for their constructive feedback, which helped improve the quality of this manuscript.

Declarations

The authors declare no conflict of interest.

Declaration of Generative AI and AI-assisted technologies

AI-assisted technology (ChatGPT) was used to support sentence restructuring and clarity improvements. The authors confirm that the underlying ideas, arguments, data analyses, and conclusions are original and were not generated by AI. All AI-assisted edits were critically reviewed and validated by the authors.

REFERENCES

[1] United Nations, "TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR SUSTAINABLE DEVELOPMENT," New York, USA, 2015.
[2] United Nations, "THE SUSTAINABLE DEVELOPMENT GOALS REPORT," 2024. Accessed: May 27, 2025. [Online]. Available: https://unstats.un.org/sdgs/report/2024/
[3] M. Angin *et al.*, "A ROBERTA APPROACH FOR AUTOMATED PROCESSING OF SUSTAINABILITY REPORTS," *Sustainability (Switzerland)*, vol. 14, no. 23, Dec. 2022, doi: https://doi.org/10.3390/su142316139
[4] G. Han, J. Tsao, and X. Huang, "LENGTH-AWARE MULTI-KERNEL TRANSFORMER FOR LONG DOCUMENT CLASSIFICATION," May 2024, [Online]. Available: http://arxiv.org/abs/2405.07052
[5] M. T. Lafleur, "USING LARGE LANGUAGE MODELS TO HELP TRAIN MACHINE LEARNING SDG CLASSIFIERS," *UN Desa Working Paper 180*, Nov. 2023, [Online]. Available: https://desapublications.un.org/working-papers.
[6] A. Hajikhani and A. Suominen, "MAPPING THE SUSTAINABLE DEVELOPMENT GOALS (SDGS) IN SCIENCE, TECHNOLOGY AND INNOVATION: APPLICATION OF MACHINE LEARNING IN SDG-ORIENTED ARTEFACT DETECTION," *Scientometrics*, vol. 127, no. 11, pp. 6661–6693, Nov. 2022, doi: https://doi.org/10.1007/s11192-022-04358-x
[7] Y. Liu *et al.*, "ROBERTA: A ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH," in *ICLR 2020 Conference Blind Submission*, Jul. 2019.
[8] C. Y. Sy, L. L. Maceda, M. Joy, P. Canon, and N. M. Flores, "BEYOND BERT: EXPLORING THE EFFICACY OF ROBERTA AND ALBERT IN SUPERVISED MULTICLASS TEXT CLASSIFICATION," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 3, pp. 45–53, 2024, doi: https://doi.org/10.14569/IJACSA.2024.0150323
[9] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "LEVERAGING LARGE LANGUAGE MODELS IN TOURISM: A COMPARATIVE STUDY OF THE LATEST GPT OMNI MODELS AND BERT NLP FOR CUSTOMER REVIEW CLASSIFICATION AND SENTIMENT ANALYSIS," *Information (Switzerland)*, vol. 15, no. 12, Dec. 2024, doi: https://doi.org/10.3390/info15120792
[10] M. Sebők, V. Kovács, M. Bánóczy, D. M. Eriksen, N. Neptune, and P. Roussille, "BEYOND TOKEN LIMITS: ASSESSING LANGUAGE MODEL PERFORMANCE ON LONG TEXT CLASSIFICATION," Sep. 2025, [Online]. Available: http://arxiv.org/abs/2509.10199
[11] D. Trautmann, "LARGE LANGUAGE MODEL PROMPT CHAINING FOR LONG LEGAL DOCUMENT CLASSIFICATION," Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.04138
[12] K. Nawab *et al.*, "FINE-TUNING FOR ACCURACY: EVALUATION OF GENERATIVE PRETRAINED TRANSFORMER (GPT) FOR AUTOMATIC ASSIGNMENT OF INTERNATIONAL CLASSIFICATION OF DISEASE (ICD) CODES TO CLINICAL DOCUMENTATION," *J Med Artif Intell*, vol. 7, no. June, 2024, doi: https://doi.org/10.21037/jmai-24-60
[13] OSDG, U. I. S. D. G. A. I. Lab, and PPMI, "OSDG COMMUNITY DATASET (OSDG-CD)," Apr. 2024, *Zenodo*.
[14] A. Kurniasih and L. Parningotan Manik, "ON THE ROLE OF TEXT PREPROCESSING IN BERT EMBEDDING-BASED DNNS FOR CLASSIFYING INFORMAL TEXTS," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022, doi: https://doi.org/10.14569/IJACSA.2022.01306109
[15] M. Siino, I. Tinnirello, and M. La Cascia, "IS TEXT PREPROCESSING STILL WORTH THE TIME? A COMPARATIVE SURVEY ON THE INFLUENCE OF POPULAR PREPROCESSING METHODS ON TRANSFORMERS AND TRADITIONAL CLASSIFIERS," *Inf Syst*, vol. 121, Mar. 2024, doi: https://doi.org/10.1016/j.is.2023.102342
[16] V. Singh, M. Pencina, A. J. Einstein, J. X. Liang, D. S. Berman, and P. Slomka, "IMPACT OF TRAIN/TEST SAMPLE REGIMEN ON PERFORMANCE ESTIMATE STABILITY OF MACHINE LEARNING IN CARDIOVASCULAR IMAGING," *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: https://doi.org/10.1038/s41598-021-93651-5
[17] T. B. Brown *et al.*, "LANGUAGE MODELS ARE FEW-SHOT LEARNERS," in *Advances in Neural Information Processing Systems*, vol. 33, Curran Associates, Inc, 2020, pp. 1877–901.
[18] M. Skrynnyk, G. Disassa, A. Krachkov, and J. Devera, "SDGI CORPUS: A COMPREHENSIVE MULTILINGUAL DATASET FOR TEXT CLASSIFICATION BY SUSTAINABLE DEVELOPMENT GOALS," in *Proceedings of the 2nd Symposium on NLP for Social Good (NSG 2024)*, United Kingdom, Apr. 2024.

[19] R. Pan, J. A. García-Díaz, and R. Valencia-García, "COMPARING FINE-TUNING, ZERO AND FEW-SHOT STRATEGIES WITH LARGE LANGUAGE MODELS IN HATE SPEECH DETECTION IN ENGLISH," *CMES - Computer Modeling in Engineering and Sciences*, vol. 140, no. 3, pp. 2849–2868, 2024, doi: https://doi.org/10.32604/cmes.2024.049631
[20] E. Latif and X. Zhai, "FINE-TUNING CHATGPT FOR AUTOMATIC SCORING," *Computers and Education: Artificial Intelligence*, vol. 6, Jun. 2024, doi: https://doi.org/10.1016/j.caeai.2024.100210
[21] Z. Khundmiri, "COMPARATIVE ANALYSIS OF ALGORITHMS IN GPT-3: A SURVEY ON PERFORMANCE, TRAINING, AND FINE-TUNING STRATEGIES," *International Journal of Engineering Research & Technology (IJERT)*, vol. 11, no. 06, 2023, [Online]. Available: www.ijert.org
[22] Y. Liu *et al.*, "ROBERTA: A ROBUSTLY OPTIMIZED BERT PRETRAINING APPROACH," *arXiv preprint*, Jul. 2019.
[23] R. Mohawesh, H. Bany Salameh, Y. Jararweh, M. Alkhalaileh, and S. Maqsood, "FAKE REVIEW DETECTION USING TRANSFORMER-BASED ENHANCED LSTM AND ROBERTA," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 250–258, Jan. 2024, doi: https://doi.org/10.1016/j.ijcce.2024.06.001
[24] A. Hussain, A. Saadia, and F. M. Alserhani, "RANSOMWARE DETECTION AND FAMILY CLASSIFICATION USING FINE-TUNED BERT AND ROBERTA MODELS," *Egyptian Informatics Journal*, vol. 30, Jun. 2025, doi: https://doi.org/10.1016/j.eij.2025.100645
[25] M. S. I. Malik, A. Nazarova, M. M. Jamjoom, and D. I. Ignatov, "MULTILINGUAL HOPE SPEECH DETECTION: A ROBUST FRAMEWORK USING TRANSFER LEARNING OF FINE-TUNING ROBERTA MODEL," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 8, Sep. 2023, doi: https://doi.org/10.1016/j.jksuci.2023.101736
[26] M. A. Talukder *et al.*, "A HYBRID DEEP LEARNING MODEL FOR SENTIMENT ANALYSIS OF COVID-19 TWEETS WITH CLASS BALANCING," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-97778-7
[27] I. Loshchilov and F. Hutter, "DECOUPLED WEIGHT DECAY REGULARIZATION," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1711.05101
[28] D. Masters and C. Luschi, "REVISITING SMALL BATCH TRAINING FOR DEEP NEURAL NETWORKS," Apr. 2018.
[29] G. K. M *et al.*, "HYBRID OPTIMIZATION DRIVEN FAKE NEWS DETECTION USING REINFORCED TRANSFORMER MODELS," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-99936-3
[30] V. Ganganwar and R. Rajalakshmi, "EMPLOYING SYNTHETIC DATA FOR ADDRESSING THE CLASS IMBALANCE IN ASPECT-BASED SENTIMENT CLASSIFICATION," *Journal of Information and Telecommunication*, vol. 8, no. 2, pp. 167–188, 2024, doi: https://doi.org/10.1080/24751839.2023.2270824
[31] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "IMPROVING LANGUAGE UNDERSTANDING BY GENERATIVE PRE-TRAINING," 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[32] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "LANGUAGE MODELS ARE UNSUPERVISED MULTITASK LEARNERS," San Francisco, CA, USA, 2019. Accessed: Oct. 12, 2025. [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[33] A. Vaswani *et al.*, "ATTENTION IS ALL YOU NEED," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6000–6010.
[34] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, and O. Isaac Abiodun, "A COMPREHENSIVE STUDY OF CHATGPT: ADVANCEMENTS, LIMITATIONS, AND ETHICAL CONSIDERATIONS IN NATURAL LANGUAGE PROCESSING AND CYBERSECURITY,", *Multidisciplinary Digital Publishing Institute (MDPI)*, Aug. 01, 2023, doi: https://doi.org/10.3390/info14080462
[35] P. P. Ray, "CHATGPT: A COMPREHENSIVE REVIEW ON BACKGROUND, APPLICATIONS, KEY CHALLENGES, BIAS, ETHICS, LIMITATIONS AND FUTURE SCOPE," *KeAi Communications Co,* Jan. 01, 2023, doi: https://doi.org/10.1016/j.iotcps.2023.04.003
[36] P. Ohm *et al.*, "FOCUSING ON FINE-TUNING: UNDERSTANDING THE FOUR PATHWAYS FOR SHAPING GENERATIVE AI," *Columbia Sci. Technol. Law Rev*, vol. 25, no. 214, pp. 214–245, 2024, [Online]. Available: https://perma.cc/YDY7-ZAS6. https://doi.org/10.52214/stlr.v25i2.12762
[37] OpenAI, "FINE-TUNING GPT MODELS." Accessed: Jun. 21, 2025. [Online]. Available: https://platform.openai.com/docs/guides/fine-tuning
[38] H. Yang, Y. Zhang, J. Xu, H. Lu, P. Ann Heng, and W. Lam, "UNVEILING THE GENERALIZATION POWER OF FINE-TUNED LARGE LANGUAGE MODELS," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, Mexico: Association for Computational Linguistics, pp. 884–899, 2024, doi: https://doi.org/10.18653/v1/2024.naacl-long.51
[39] T.-G. Marchitan, C. Creanga, and L. P. Dinu, "TEAM UNIBUC-NLP AT SEMEVAL-2024 TASK 8: TRANSFORMER AND HYBRID DEEP LEARNING BASED MODELS FOR MACHINE-GENERATED TEXT DETECTION," in *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Mexico: Association for Computational Linguistics, pp. 403–411, Jun. 2024, doi: https://doi.org/10.18653/v1/2024.semeval-1.63
[40] J. Howard and S. Ruder, *Universal Language Model Fine-tuning for Text Classification*. 2018, doi: https://doi.org/10.18653/v1/P18-1031
[41] S. Casola, I. Lauriola, and A. Lavelli, "PRE-TRAINED TRANSFORMERS: AN EMPIRICAL COMPARISON," *Machine Learning with Applications*, vol. 9, p. 100334, Sep. 2022, doi: https://doi.org/10.1016/j.mlwa.2022.100334
[42] M. K. S. Ma'aitah, A. Helwan, and A. Radwan, "URINARY BLADDER ACUTE INFLAMMATIONS AND NEPHRITIS OF THE RENAL PELVIS: DIAGNOSIS USING FINE-TUNED LARGE LANGUAGE MODELS," *J Pers Med*, vol. 15, no. 2, Feb. 2025, doi: https://doi.org/10.3390/jpm15020045
[43] W. K. S. Ojemann *et al.*, "ZERO-SHOT EXTRACTION OF SEIZURE OUTCOMES FROM CLINICAL NOTES USING GENERATIVE PRETRAINED TRANSFORMERS," *J Healthc Inform Res*, Sep. 2025, doi: https://doi.org/10.1007/s41666-025-00198-5
[44] OpenAI, "GPT-4 Technical Report," 2024.
[45] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A COMPREHENSIVE EVALUATION OF OVERSAMPLING TECHNIQUES FOR ENHANCING TEXT CLASSIFICATION PERFORMANCE," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: https://doi.org/10.1038/s41598-025-05791-7

[46] I. Tabassum and V. Nunavath, "A HYBRID DEEP LEARNING APPROACH FOR MULTI-CLASS CYBERBULLYING CLASSIFICATION USING MULTI-MODAL SOCIAL MEDIA DATA," *Applied Sciences (Switzerland)*, vol. 14, no. 24, Dec. 2024, doi: https://doi.org/10.3390/app142412007

[47] M. Atay, "A COMPARATIVE STUDY OF PROMPTING AND FINE-TUNING FOR BINARY TEXT CLASSIFICATION OF SUSTAINABLE DEVELOPMENT GOALS," Middle East Technical University, Turkey, 2024.

[48] M. Sushil *et al.*, "A COMPARATIVE STUDY OF LARGE LANGUAGE MODEL-BASED ZERO-SHOT INFERENCE AND TASK-SPECIFIC SUPERVISED CLASSIFICATION OF BREAST CANCER PATHOLOGY REPORTS," *Journal of the American Medical Informatics Association*, vol. 31, no. 10, pp. 2315–2327, Oct. 2024, doi: https://doi.org/10.1093/jamia/ocae146