Leveraging Large Language Models to Build Computationally Efficient Models for Sustainable Finance Investment Decision Support

Loris Bergeron
SnT - SEDAN
Banque de Luxembourg
Luxembourg, Luxembourg
loris.bergeron@blu.bank

Jérôme François
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
jerome.francois@uni.lu

Radu State
SnT - SEDAN
University of Luxembourg
Luxembourg, Luxembourg
radu.state@uni.lu

Jean Hilger
SnT - Finnovation Hub
University of Luxembourg
Luxembourg, Luxembourg
jean.hilger@uni.lu

***Abstract*—Assessing companies' contributions to the United Nations Sustainable Development Goals (SDGs) is essential for sustainable investment and regulatory reporting. However, extracting reliable insights from heterogeneous textual sources remains a challenge due to limited labeled data, domain imbalance, and privacy constraints. We present LëtzSDG, a lightweight BERT-based multiclass classifier fine-tuned on a hybrid dataset constructed using Large Language Models (LLMs). Multiple LLMs are used to (i) expand domain-specific SDG keywords, (ii) perform consensus-based zero-shot labeling, and (iii) generate synthetic data to balance underrepresented classes. Unlike cloud-hosted LLMs, LëtzSDG is designed for on-premises deployment within financial institutions, ensuring data-privacy compliance. Integrated into a human-in-the-loop investment workflow, its predictions are span-linked for traceability and committee review. Evaluated on public datasets (the OSDG Community Dataset and the SDG Integration Corpus), LëtzSDG outperforms SDG-specific baselines, a strong NLI-based zero-shot model, and several open LLMs, while rivaling larger closed models at a fraction of their size. LëtzSDG and its datasets are publicly available on Hugging Face.**

***Index Terms*—Sustainable Development Goals (SDGs), Sustainable Finance, Text Classification, Large Language Models (LLMs), Human-in-the-Loop, On-Premises Deployment**

I. Introduction

The financial sector plays an essential role in the achievement of the Sustainable Development Goals (SDGs) [1] by directing investments toward companies that contribute to ecological and social transitions (Table I).

Specifically, the EU Sustainable Finance Disclosure Regulation (SFDR) [2] requires disclosure of non-financial criteria to encourage transparent sustainable investments. However, concerns remain about the credibility of these disclosures [3].

In addition to the regulatory landscape, the financial sector faces an increasing number of obligations. In practice, investment managers spend a significant part of their time on manual tasks to comply with these obligations [4]. These activities distract them from their core competencies: portfolio optimization and day-to-day asset management.

In fact, evaluating contributions to the SDGs is challenging, as the goals mix quantitative and qualitative targets, making consistent extraction of relevant information from text difficult. Thus, companies such as Clarity AI[1] and SESAMm[2], provide SDG-related insights using proprietary NLP models. However, these platforms operate as closed, cloud-based services, raising concerns for financial institutions about data-privacy, explainability, and regulatory compliance.

In this paper, we introduce LëtzSDG, a lightweight BERT-based multiclass classifier[3] [5] designed to identify SDG-related insights in textual data. As a result, LëtzSDG is suitable for on-premises deployment due to its limited size. Thus, it integrates easily into investment workflow that involves human oversight, enabling investment managers to assess companies' contributions to the SDGs using internal or public documents without exposing the latter to an external third-party service.

A key challenge in building such a classifier is the scarcity of labeled training data. To address this, we design a data-centric pipeline leveraging Large Language Models (LLMs) to build a hybrid dataset that combines real-world texts with synthetic samples. This approach improves the diversity of training while preserving the precision of fine-tuned classifiers.

TABLE I: The 17 Sustainable Development Goals

SDG	Description
SDG1	No Poverty
SDG2	Zero Hunger
SDG3	Good Health and Well-being
SDG4	Quality Education
SDG5	Gender Equality
SDG6	Clean Water and Sanitation
SDG7	Affordable and Clean Energy
SDG8	Decent Work and Economic Growth
SDG9	Industry, Innovation and Infrastructure
SDG10	Reduced Inequalities
SDG11	Sustainable Cities and Communities
SDG12	Responsible Consumption and Production
SDG13	Climate Action
SDG14	Life Below Water
SDG15	Life on Land
SDG16	Peace, Justice and Strong Institutions
SDG17	Partnerships for the Goals

[1]https://clarity.ai/one-platform
[2]https://www.sesamm.com/esg-insights
[3]https://hf.co/google-bert/bert-base-uncased

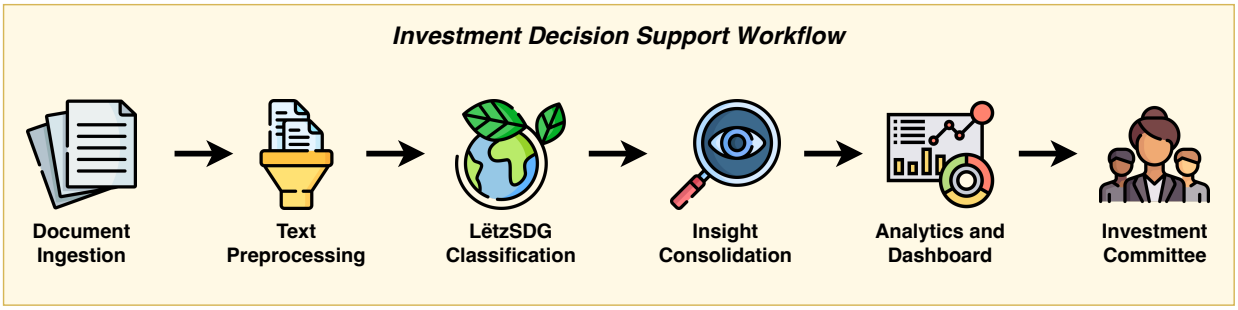


Fig. 1: **Investment Decision Workflow.** Overview of the human-in-the-loop process enabled by LëtzSDG.

Our main contributions are thus fourfold:

- A data-centric pipeline to build a hybrid training dataset that leverages multiple LLMs to combine real-world and synthetic samples for SDG classification tasks.
- LëtzSDG, a multiclass BERT-based classifier that can be integrated on-premises into human-in-the-loop investment workflows. The full method is presented in this paper to promote transparency in line with the EU AI Act. Both the model and the hybrid training dataset are publicly available on Hugging Face[4].
- A transparent and reproducible evaluation on two public datasets, the OSDG Community Dataset (OSDG-CD) [6] and the SDG Integration Corpus (SDGi Corpus) [7], demonstrating a strong generalization to unseen data.
- A comparative analysis of our model against SDG-specific solutions, a strong NLI-based zero-shot baseline, and open or closed-source general-purpose LLMs.

Developed in collaboration with Banque de Luxembourg[5], LëtzSDG is tailored for real-world deployment in highly regulated financial environments. One of the co-authors, affiliated with the bank through an academic-industry partnership, worked closely with domain experts from the same institution to ensure regulatory alignment and practical relevance.

The paper is structured as follows. Section II describes the industrial use case. Section III details LëtzSDG. Section IV presents our experimental setup, and Section V reports the comparative results. We conclude the paper in Section VI.

II. Case Study Example

Although LëtzSDG is not specific to any SDG, this section illustrates the workflow on SDG 5 (Gender Equality). Assessing it relies on sensitive internal documents-like HR policies, board minutes, and due diligence reports-which cannot be processed via external cloud services due to confidentiality.

To address these limitations, LëtzSDG is built for on-premises deployment and supports an end-to-end investment decision workflow. As highlighted in Fig. 1, during *document ingestion* internal and public documents are uploaded followed by *text preprocessing* to clean and segment the content. Then, LëtzSDG can identify text extracts related to SDG 5, such as mentions of equal pay, or leadership diversity.

[4]https://hf.co/collections/lrsbrgrn/letzsdg

[5]https://banquedeluxembourg.com

Once done, *Insight consolidation* aggregates these classifications at the document and company levels, linking them back to their original text sources to maintain traceability. This enables structured comparisons between companies and reports. These insights are then exposed through an *analytics and dashboard* layer, allowing investment managers to examine SDG coverage and trends over time or portfolios. Finally, the *investment committee* reviews the results in the form of structured traceable evidence. These insights inform investment managers' decisions by clearly indicating how company practices are related to SDG 5. Consequently, all model outputs are reviewed by the investment committee.

Thus, our workflow aligns with the observation of the European Securities and Markets Authority (ESMA) [3] that SDG funds often require the assembly of information from multiple textual sources, making the assessment challenging. LëtzSDG bridges this gap by converting internal or public documents into structured, traceable, and auditable evidence, with human oversight, while remaining on-premises to comply with financial-sector requirements.

III. Method

A. LëtzSDG Overview

LëtzSDG is a multiclass BERT model fine-tuned for SDG classification. It classifies text into one of the 17 SDGs. BERT was chosen for its efficiency and real-world suitability. Fine-tuned Small Language Models (SLMs) often outperform general-purpose LLMs in domain-specific tasks [8], making BERT well-suited for low-resource computing environments.

The performance of LLMs is closely linked to both the size and the quality of the training data [9]. This paper focuses on a data-centric pipeline to build a relevant dataset before training a task-specific classifier, as shown in Fig. 2.

Thus, we built a hybrid training dataset starting with data from the FineWeb dataset [10]. To extract SDG-relevant content, we applied keyword filters from the University of Auckland (UoA) [11] and the Joint Research Centre (JRC) [12], extended with LLM-generated synonyms. This leads to a set of 17 filters, one per SDG, to extract up to 3,000 texts per goal, yielding 51,000 real samples. However, to improve reliability, the extracts were annotated by multiple LLMs. For a given extract, if and only if they all agree on the same SDG, it is kept to minimize individual model biases.

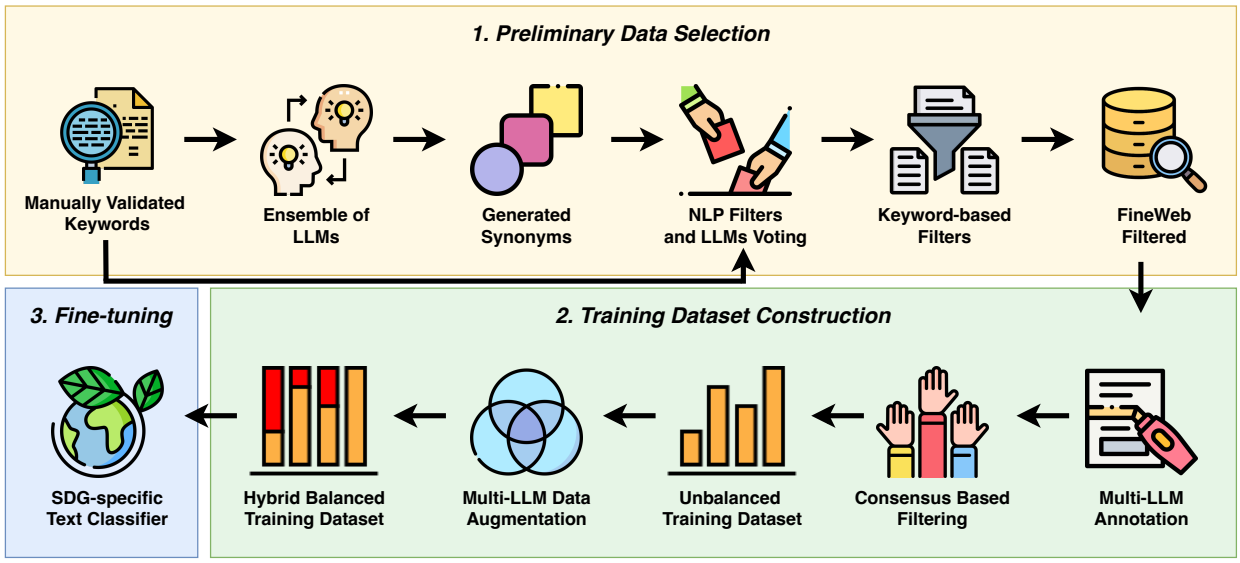


Fig. 2: **Hybrid Dataset Construction Pipeline.** Workflow for building the hybrid dataset and fine-tuning LëtzSDG.

However, the samples are imbalanced across SDGs. To mitigate this, the same LLMs generate synthetic samples via data augmentation, varying tone, length, and style. The final hybrid dataset, which combines real and synthetic texts, is finally used to fine-tune BERT for the SDG classification.

Finally, LëtzSDG is designed to operate from any textual source (internal or public), such as regulatory disclosures, company reports, and internal HR documents; extending to tables and figures is deferred to future multimodal versions. We emphasize that this is an explicit design choice to first validate a robust text-only classifier.

Importantly, LëtzSDG is not an automated compliance verification system, it surfaces SDG-related spans that remain linked to their source documents and are reviewed by the investment committee, ensuring auditable, governance-aligned workflows that enhance transparency.

B. Preliminary Data Selection

As highlighted earlier, we first need to build a base of relevant text extracts by searching real-world data using keywords.

Keyword-based Filters. Manually validated keywords (MVKs) often lack vocabulary diversity due to human bias. To enhance coverage, we used ten distinct LLMs to generate synonyms [13]. Given an MVK k associated with an SDG from UoA or JRC, an LLM a generates a synonym $syn_s^a(k)$ (Appendix A). Together, they form the new set of keywords for a given SDG.

Formally, let $LLMs$ denote the set of all LLMs used, this results in SDG_i^a as the set of keywords for SDG i, with $a \in \text{MVK}, \text{LLMs}$ and $i \in 1, \ldots, 17$. To ensure validity, only keywords generated by all LLMs are retained. The final set of LLM-generated terms for SDG i is SDG_i^{LLMs}.

Then, the complete keyword set is defined as:

$$SDG_i^{LetzSDG} = SDG_i^{MVK} \cup SDG_i^{LLMs} \tag{1}$$

This forms the 17 keyword-based filters to accurately keep relevant SDG-related information in a set of texts. In addition, NLP filters (e.g., removing duplicates, converting plurals to singulars) are applied.

FineWeb as the Foundation. Our keyword filters are applied to FineWeb [10], a high-quality and large-scale open dataset created by crawling web sources.

Specifically, we used the 10TB sample[6], denoted D_0, and applied a length filter f_1 to meet BERT's 512-token limit:

$$D = \{x \in D_0 \mid f_1(x)\} \tag{2}$$

As FineWeb is domain-agnostic, our 17 keywords allow to extract SDG-relevant samples. For each SDG_i, we collect a result set r_i and randomly sample 3,000 items, defined as:

$$D' = \bigcup_{i=1}^{17} r_i, \quad |D'| = 51{,}000 \tag{3}$$

C. Training Dataset Construction

Multi-LLM Consensus Annotation. While keyword filtering identifies SDG-relevant segments, it lacks precision in labeling [13]. To improve label quality without human annotation [14], we use a consensus strategy across an ensemble of three LLMs: Qwen2.5-32B [15], Gemma 3-27B [16], and Mistral Small 3.1-24B [17]. Each model predicts SDG labels in a Zero-Shot (ZS) setting [18], using a structured JSON prompt [19] (Appendix C). ZS was adopted to take advantage of pre-trained knowledge of the models, eliminating the need for curated examples.

We define the consensus as $D'_i = \{x \in D' \mid f_1(x) = f_2(x) = f_3(x) = i\}$, where $f_k(x)$ is the label from LLM k. Thus, only samples with full agreement are retained for SDG i.

[6]https://hf.co/datasets/HuggingFaceFW/fineweb

The resulting dataset is unbalanced, with SDG 3 having 4,157 samples and SDG 17 only 92 while the class imbalance is known to degrade BERT performance and will therefore have a negative impact on LëtzSDG [20].

Multi-LLM Data Augmentation. To address the problem mentioned above without an excessive generation cost [21], we combine raw and synthetic data using Data Reformation (DR) [22]. An ensemble of LLMs (Mistral Small 3.1-24B, Gemma 3-27B, Qwen2.5-32B) reformulate samples into more diverse forms [23], [24].

With SDG 3 as the majority class ($|D'_3| = 4{,}157$), the number of synthetic samples needed for each class SDG_i is:

$$m(SDG_i) = |D'_3| - |D'_i|, i = 1, 2, 4, \ldots, 17 \quad (4)$$

We select $m(SDG_i)$ existing samples for each D'_i except D'_3 and each is reformed using a random prompt (Appendix B) and a random temperature in [0.2, 0.7]. Each LLM generates one third of the synthetic samples, capturing variation in tone, style, and length [25], [26]. The final balanced dataset D_{final} combines raw and synthetic samples in variable proportions across SDGs as shown in Fig. 3.

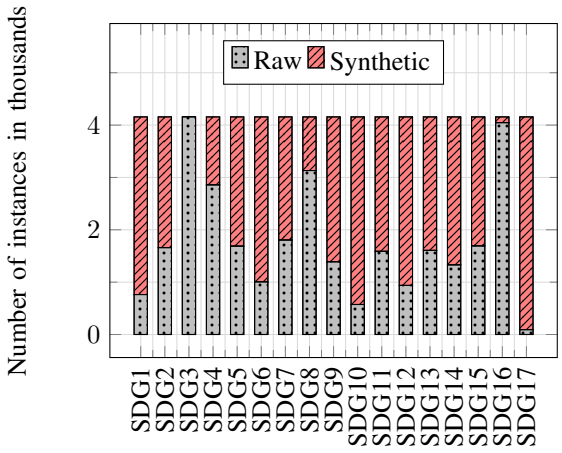


Fig. 3: Data distribution in the training dataset

D. Fine-tuning

Using the hybrid dataset, a model can be fine-tuned to classify new text extracts among the SDGs. To achieve this, we perform standard supervised fine-tuning. We do not distill model weights from a teacher model; our BERT-based classifier is an SLM fine-tuned for our specific task to reach performance of the LLM ensemble used for label consensus annotation. Moreover, previous work shows that LLM-generated labels can rival human quality [27].

The model is fine-tuned on 3 epochs (batch size: 16). Mixed-precision (bfloat16) reduced memory usage. Optimization used AdamW (lr: 2×10^{-5}, weight decay: 0.01) and 10% warm-up. Inputs were truncated to 512 tokens, and evaluation was performed after each epoch. Full training parameters are provided in the Appendix D.

IV. Experimental Setup

Validation Set. To assess generalization and ensure reproducibility, we use two public datasets solely for evaluation: OSDG Community Dataset (OSDG-CD) [6] and SDG Integration Corpus (SDGi Corpus) [7]. Thus, all reported experiments use exclusively public datasets; therefore, no confidential or internal documents were used.

OSDG-CD includes 43,010 English excerpts, each labeled with a single SDG. We randomly sampled 1,000 texts with an agreement_score ≥ 0.75. Since SDG 17 is missing from OSDG-CD, we expanded the validation set with 869 additional samples from the SDGi Corpus but only English, single-label texts with lengths below 512 tokens.

The resulting validation set contains 1,869 samples. The 17 SDGs are represented but in an unbalanced manner (e.g., $n(\text{SDG } 5) = 183$, $n(\text{SDG } 17) = 54$; Fig. 4).

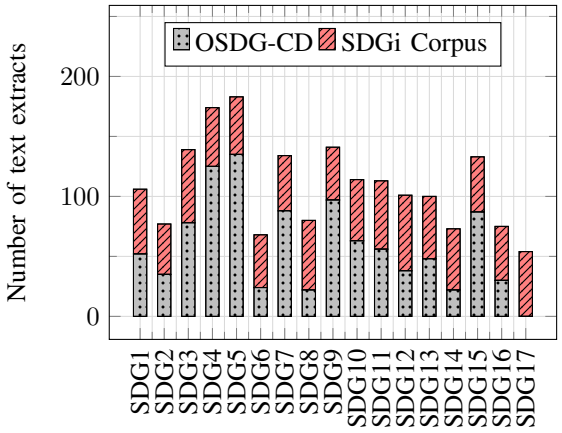


Fig. 4: Distribution of classes in the validation set

Baseline Methods. We compared LëtzSDG with SDG-specific solutions, NLI-based ZS classifier, and general-purpose LLMs. Among SDG-specific solutions, SDG Mapper [12] is keyword-based, while Aurora [28] and Elsevier [29] use mBERT.

To foster a fair comparison, we include BART-Large-MNLI[7] as a strong ZS NLI baseline. We use the 17 SDG class codes (e.g., SDG_1_NO_POVERTY) as candidate labels and select the label with the highest entailment score.

For general-purpose LLMs, we tested five open models (Llama3.2-3B, Llama3.1-8B [30], Mistral Small 3.1-24B [31], Gemma 3-27B [16], Qwen2.5-32B [32]) and four closed models (Claude 3.5 Haiku [33], Mistral Large 2 [17], GPT-3.5 [34], GPT-4o [35]).

All LLMs were evaluated in the ZS setting to align with the consensus-labeling step using prompt in the Appendix C. We excluded ZS-CoT [36] due to its limited impact on smaller models and its higher inference cost [37].

Hardware and Energy Profile. We trained LëtzSDG on a T4 GPU (70W TDP) for 2.5 hours, consuming 0.17 kWh and emitting an estimated 0.05 kgCO_2eq, as calculated by the Machine Learning Impact Calculator (MLIC) [38].

[7]https://hf.co/facebook/bart-large-mnli

V. Results

A. Model Evaluation

Performance Comparison. Table II reports macro-F1 scores for SDG-specific solutions, ZS-NLI classifier, open and closed LLMs, and LëtzSDG. To ensure a fair comparison, we rely on the SDG with the highest prediction from the multi-label classifiers (i.e., SDG Mapper and Elsevier), aligning with our multiclass validation setup.

LëtzSDG outperforms all SDG-specific solutions, with gains of up to +32 F1 points. Regarding the ZS-NLI classifier, LëtzSDG also performs better (+29). Among open LLMs, models such as Mistral Small (24B), Gemma 3 (27B), and Qwen2.5 (32B) achieve top scores, while others such as Llama 3.2 (3B) and Llama 3.1 (8B) perform moderately. With only 110M parameters, LëtzSDG matches or exceeds several open LLMs and falls just 2–3 points behind the most accurate ones.

Closed LLMs (GPT-3.5, GPT-4o, Claude 3.5 Haiku, Mistral Large 2) achieve the highest scores (0.88–0.91), but LëtzSDG remains only 3-6 points behind. However, benchmark contamination may affect the results of closed models [39].

In general, LëtzSDG performs strongly, highlighting the potential of compact, task-specialized SLMs to rival general-purpose LLMs. Furthermore, fine-tuning LëtzSDG on keyword-annotated data yields an F1 macro of 0.64 (-0.21) while using only raw data achieves 0.81 (-0.04). On the minority SDG 17 class, F1-macro improved from 0.29 (raw) to 0.72 (hybrid), demonstrating that balancing training dataset strongly benefits the underrepresented classes.

TABLE II: Macro F1-scores across all evaluated methods

Type	Method	Params	F1-score
SDG Specific	SDG Mapper	NA	0.65
	Aurora	168M	0.59
	Elsevier	168M	0.53
ZS-NLI	BART-Large-MNLI	407M	0.56
Open LLMs	Llama 3.2	3B	0.68
	Llama 3.1	8B	0.81
	Mistral Small 3.1	24B	0.84
	Gemma 3	27B	0.88
	Qwen 2.5	32B	0.87
Closed LLMs	Claude 3.5 Haiku	-	0.88
	Mistral Large 2	123B	0.88
	GPT-3.5	175B	0.88
	GPT-4o	-	0.91
Our	LëtzSDG	110M	0.85

Efficiency Evaluation. We define efficiency as the ratio between the divided F1-score and the size of the model (in billions of parameters). SDG Mapper is excluded due to its non-LLM, keyword-based approach; Claude 3.5 Haiku and GPT-4o are omitted as their sizes are unknown.

LëtzSDG achieves an efficiency score of 7.73, outperforming all other approaches. Among SDG-specific solutions, Aurora (3.51) and Elsevier (3.15) achieve less than half of this performance. Regarding BART-Large-MNLI (1.38), its efficiency is nearly six times lower than that of LëtzSDG.

General-purpose LLMs perform significantly worse: Llama 3.2 (3B) and 3.1 (8B) score just 0.23 and 0.10. Larger models like Mistral Small 3.1 (24B), Gemma 3 (27B), and Qwen2.5 (32B) yield F1s of 0.84–0.88, but with efficiencies near 0.03. Mistral Large (123B) and GPT-3.5 (175B) drop to just 0.005.

As shown in Fig. 5, the largest models provide only marginal performance gains. In contrast, LëtzSDG achieves near-top-tier accuracy with just 110M parameters, combining efficiency and performance through design rather than scale. For example, on a MacBook M3, LëtzSDG processed the evaluation dataset (1,869 texts) in 66.5 seconds, about 0.036 seconds per sample, demonstrating that it can operate efficiently even in low-resource computing environments.

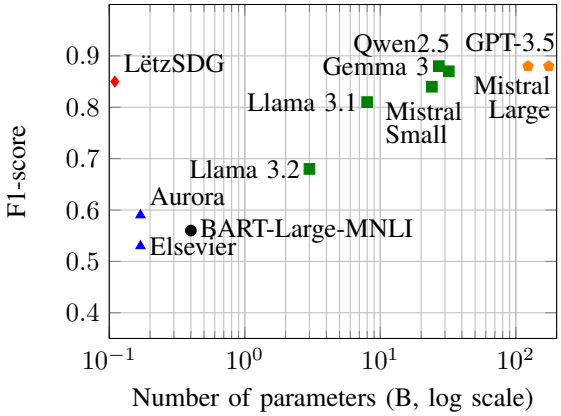


Fig. 5: Effect of model size on performance

Variations Across SDGs. We evaluate the performance variability of LëtzSDG across the SDGs. As shown in Fig. 6, LëtzSDG shows consistent strong performance across a wide range of SDGs. In contrast, SDG-specific solutions, ZS-NLI (Fig. 6a) and open LLMs (Fig. 6b) exhibit greater variability.

LëtzSDG performs competitively with closed LLMs (Fig. 6c) except for SDG 17.

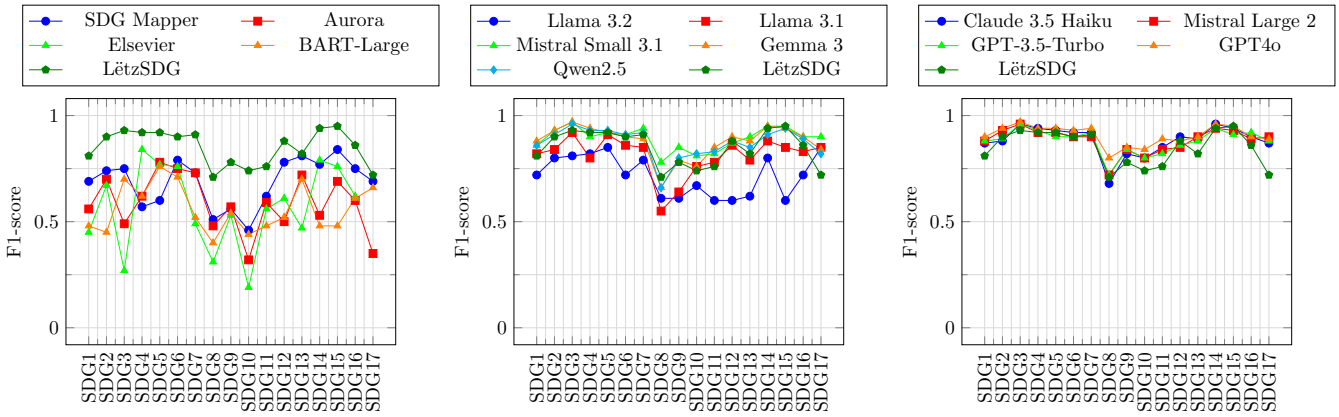
In general, LëtzSDG offers a robust, balanced alternative, outperforming SDG-specific solutions in performance and consistency. While closed LLMs may excel on some SDGs, they come with higher usage costs and larger model sizes.

B. Hybrid Dataset Evaluation

The hybrid dataset construction is the core component of LëtzSDG. This section validates its two key steps: SDG labeling of real texts and generation of synthetic samples.

Human-LLM Annotation Alignment. First, we evaluated if the annotated labels by LLM align with human judgments. We sampled 170 examples across the 17 SDGs (10 per goal) and split them into two subsets: a consensus subset, where all three LLMs agreed on the SDG label, and a non-consensus subset, where only one model assigned a label and the others disagreed. Two human annotators independently reviewed a random, non-overlapping set of examples, giving a Yes or No judgment on SDG correctness for the input text.

Human agreement with LLM-annotated labels was measured as the proportion of examples judged correct. Agreement was significantly higher in the consensus subset (88.24%) compared to the non-consensus subset (41.18%), indicating that multi-LLM consensus improves the reliability of the label.



(a) vs. SDG-specific & ZS-NLI (b) vs. Open LLMs (c) vs. Closed LLMs

Fig. 6: F1-score variations across the 17 SDGs comparing SDG-specific and ZS-NLI baselines, open and closed LLMs.

Quality Analysis of Synthetic Data. To assess the quality of synthetic data relative to their corresponding raw examples, we applied G-Eval [40], which evaluates Relevance (Rel), Coherence (Coh), and Consistency (Con) (each on a 1-5 scale), and Fluency (Flu) (on a 1-3 scale). Unlike traditional metrics such as ROUGE [41], G-Eval has demonstrated significantly stronger correlation with human judgment [42].

SDG 3 was excluded as no synthetic data was generated for it. We observed consistently high quality in the synthetic data. Relevance, Coherence, and Consistency scores are tightly clustered near the upper bounds, reflecting contextual alignment with their raw data, logical structure, and factual consistency. Fluency scores are uniformly high, confirming that the synthetic texts are clear and well-formed (Table III).

However, SDG 8 falls below the top tier, contributing to the downstream performance trends observed in Section V-A, and identifying this goal as a candidate for further refinement.

TABLE III: Synthetic data analysis using G-Eval

SDGs	Rel	Coh	Con	Flu	Avg.
SDG1	4.35	4.23	4.55	2.97	4.02
SDG2	4.31	4.23	4.34	2.98	3.96
SDG3	-	-	-	-	-
SDG4	4.28	4.28	4.52	3.00	4.02
SDG5	4.40	4.33	4.51	2.99	4.06
SDG6	4.37	4.28	4.55	2.97	4.04
SDG7	4.39	4.27	4.51	2.99	4.04
SDG8	4.17	4.11	4.27	2.97	3.88
SDG9	4.29	4.29	4.41	2.97	3.99
SDG10	4.37	4.23	4.51	2.98	4.02
SDG11	4.20	4.23	4.26	2.98	3.92
SDG12	4.43	4.34	4.51	2.98	4.07
SDG13	4.40	4.29	4.45	2.97	4.03
SDG14	4.39	4.30	4.55	2.96	4.05
SDG15	4.31	4.23	4.51	2.96	4.00
SDG16	4.09	4.00	4.27	2.91	3.82
SDG17	4.29	4.26	4.49	2.97	4.00

Evaluating Semantic Coherence. We also semantically compared raw and synthetic samples using embeddings from nomic-embed-text-v1.5[8] and measuring cosine similarity.

[8]https://hf.co/nomic-ai/nomic-embed-text-v1.5

The raw texts showed clearer separation (0.45–0.53) compared to the synthetic texts (0.49–0.57), suggesting weaker boundaries between classes. We then applied t-SNE to assess semantic consistency within each SDG class. We observed that synthetic texts formed tighter and more coherent clusters within each class than raw texts (Fig. 7).

These results indicate our hybrid generation pipeline slightly reduces inter-class semantic separation but improves intra-class coherence. This makes the hybrid dataset suitable for SDG-related tasks that require semantically consistent data.

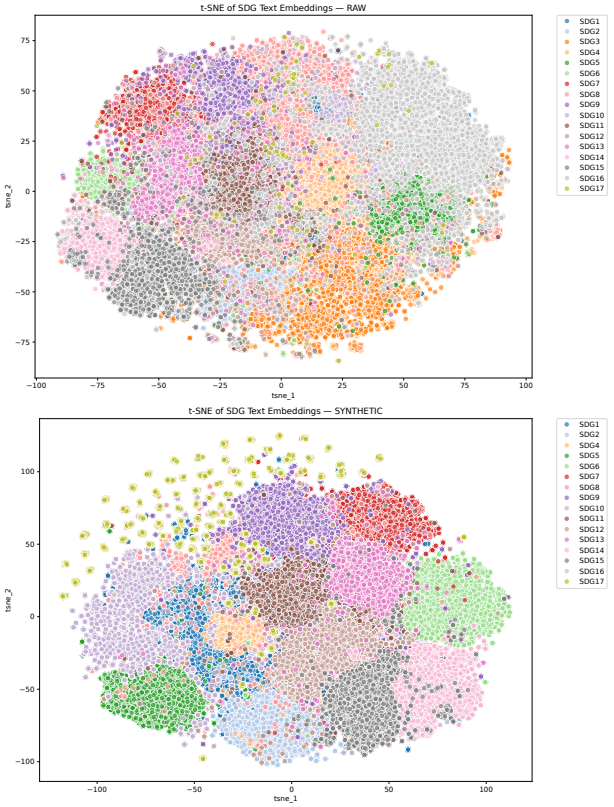


Fig. 7: t-SNE projections of text embeddings for raw data (top) and synthetic data (bottom).

VI. Conclusion

We presented LëtzSDG, a 110M-parameter multiclass BERT-based model for SDG text classification, fine-tuned through a data-centric pipeline that leverages LLMs for domain keyword expansion, multi-LLM consensus labeling, and data reformation. On a public validation set, LëtzSDG outperforms SDG-specific solutions and a strong zero-shot NLI baseline. It matches or surpasses several open LLMs and rivals much larger closed models, achieving comparable accuracy at a fraction of their size.

Our approach demonstrates the value of using LLMs for data creation rather than inference, leveraging them to build high-quality, domain-specific datasets that enable lightweight models to achieve near-LLM performance in low-resource environments. Because inference runs locally and predictions are span-linked, LëtzSDG fits on-premises human-in-the-loop finance workflows, providing auditable traceability aligned with regulatory and reporting standards.

Beyond SDGs, this workflow offers a reproducible blueprint for financial NLP: curate and balance domain data using LLMs, then deploy lightweight, auditable models that integrate seamlessly into investment decision workflows while meeting operational and data-privacy requirements.

VII. Future Work

Our next steps target four priorities: (i) we will rebalance the hybrid dataset, improving the ratio of synthetic texts, especially for low-resource goals such as SDG 17. In addition, (ii) we aim to sharpen class boundaries by adding stronger semantic filters that reduce overlaps between SDG categories, and (iii) we plan to expand LëtzSDG beyond the English text, introducing multilingual models such as mBERT[9].

Finally, we will (iv) extend to multimodal evidence (tables, figures, charts) via a lightweight OCR and table parser.

Limitations

Bias and Fairness. The hybrid training dataset is based on labels derived from a multi-LLM consensus approach. Although consensus improves consistency, it can also introduce biases present in the underlying language models, particularly for culturally nuanced SDGs. Bias audits (e.g., comparing prediction distributions across sectors) should be addressed.

Human Feedback and Adaptation. Although experts were included during development, the workflow currently lacks a reinforcement learning from human feedback (RLHF) loop. This could help align the model predictions with evolving regulatory requirements and internal investment strategies.

Reproducibility and Data Access. While evaluated on public datasets, it goes without saying that confidential internal documents cannot be released. To improve reproducibility, we provide prompts, hyperparameters, and our evaluation dataset.

Governance and Model Lifecycle. As the SFDR regulation and AI Act evolve, long-term model governance, including continuous monitoring, calibration and retraining, remains beyond the scope of this study.

[9]https://hf.co/google-bert/bert-base-multilingual-uncased

References

[1] D. o. E. United Nations and S. A.-S. Development, "Transforming our world: the 2030 Agenda for Sustainable Development," 2015, iSBN: A/RES/70/1 Pages: 16301 Type: General Assembly. [Online]. Available: https://sdgs.un.org/2030agenda

[2] European Parliament and Council, "Regulation (eu) 2019/2088 on sustainability-related disclosures in the financial services sector," November 2019. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2019/2088/oj

[3] European Securities and Markets Authority, "Trv article: Impact investing – do sdg funds fulfil their promises?" European Securities and Markets Authority, Tech. Rep. ESMA50-524821-3098, 1 2024, risk monitoring, Sustainable finance. [Online]. Available: https://www.esma.europa.eu/databases-library/esma-library

[4] European Securities and Markets Authority (ESMA), "Discussion paper on the integrated collection of funds' data," ESMA, Paris, France, Tech. Rep. ESMA12-2121844265-4904, June 2025.

[5] J. Devlin, M.-W. Chang, and K. Lee, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019, arXiv:1810.04805 [cs]. [Online]. Available: http://arxiv.org/abs/1810.04805

[6] OSDG, U. I. S. A. Lab, and PPMI, "OSDG Community Dataset (OSDG-CD)," Apr. 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.11441197

[7] M. Skrynnyk, G. Disassa, A. Krachkov, and J. DeVera, "Sdgi corpus: A comprehensive multilingual dataset for text classification by sustainable development goals," 2024.

[8] M. J. J. Bucher and M. Martini, "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification," Jun. 2024, arXiv:2406.08660 [cs]. [Online]. Available: http://arxiv.org/abs/2406.08660

[9] S. Gunasekar, Y. Zhang, and J. Aneja, "Textbooks are all you need," 2023. [Online]. Available: https://arxiv.org/abs/2306.11644

[10] G. Penedo, H. Kydlíček, and L. B. allal, "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale," Jun. 2024, arXiv:2406.17557 [cs]. [Online]. Available: http://arxiv.org/abs/2406.17557

[11] W. Wang, W. Kang, and J. Mu, "Mapping Research to the Sustainable Development Goals: A Contextualised Approach," In Review, preprint, Mar. 2023. [Online]. Available: https://www.researchsquare.com/article/rs-2544385/v3

[12] European Commission, *Mapping EU policies with the 2030 agenda and SDGs: fostering policy coherence through text based SDG mapping.* LU: Publications Office, 2023. [Online]. Available: https://data.europa.eu/doi/10.2760/110687

[13] L. Bergeron, J. François, and R. State, "ULA, a Bibliometric Method to Identify Sustainable Development Goals using Large Language Models," in *2023 IEEE International Humanitarian Technology Conference (IHTC)*. Santa Marta, Colombia: IEEE, Nov. 2023, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10508849/

[14] M. Yuan, J. Chen, Z. Xing, G. Mohammadi, and A. Quigley, "A case study of scalable content annotation using multi-llm consensus and human review," *arXiv preprint arXiv:2503.17620*, 2025.

[15] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[16] G. Team, A. Kamath, J. Ferret, and S. Pathak, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[17] Mistral AI Team, "Large enough: Introducing mistral large 2," https://mistral.ai/news/mistral-large-2407, Jul. 2024, accessed: 2025-06-12.

[18] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning – a comprehensive evaluation of the good, the bad and the ugly," 2020. [Online]. Available: https://arxiv.org/abs/1707.00600

[19] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan, "Does prompt formatting have any impact on llm performance?" 2024. [Online]. Available: https://arxiv.org/abs/2411.10541

[20] L. Mahmoudi and M. Salem, "BalBERT: A New Approach to Improving Dataset Balancing for Text Classification," *Revue d'Intelligence Artificielle*, vol. 37, no. 2, pp. 425–431, Apr. 2023. [Online]. Available: https://iieta.org/journals/ria/paper/10.18280/ria.370219

[21] E. Dauber and S. Dendekuri, "Data generation for nlp classification dataset augmentation: Using existing llms to improve dataset quality," Stanford University, Department of Computer Science, Tech. Rep., 2024.

[22] B. Ding, C. Qin, and R. Zhao, "Data Augmentation using Large Language Models: Data Perspectives, Learning Paradigms and Challenges," Jul. 2024, arXiv:2403.02990 [cs]. [Online]. Available: http://arxiv.org/abs/2403.02990
[23] Y.-S. Lee, M. Sultan, Y. El-Kurdi, T. Naseem, A. Munawar, R. Florian, S. Roukos, and R. Astudillo, "Ensemble-instruct: Instruction tuning data generation with a heterogeneous mixture of LMs," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 12 561–12 571. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.836/
[24] M. Nadas, L. Diosan, and A. Tomescu, "Synthetic data generation using large language models: Advances in text and code," 2025. [Online]. Available: https://arxiv.org/abs/2503.14023
[25] V. Veselovsky, M. H. Ribeiro, and A. Arora, "Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science," May 2023, arXiv:2305.15041 [cs]. [Online]. Available: http://arxiv.org/abs/2305.15041
[26] L. Long, R. Wang, and R. Xiao, "On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey," Jun. 2024, arXiv:2406.15126 [cs]. [Online]. Available: http://arxiv.org/abs/2406.15126
[27] N. Pangakis and S. Wolken, "Knowledge Distillation in Automated Annotation: Supervised Text Classification with LLM-Generated Training Labels," Jun. 2024, arXiv:2406.17633 [cs]. [Online]. Available: http://arxiv.org/abs/2406.17633
[28] M. Vanderfeesten, R. Jaworek, and L. Keßler, "AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs)," Zenodo, Tech. Rep., Mar. 2022, version Number: 1.0. [Online]. Available: https://zenodo.org/record/5603019
[29] R. Jaworek, "SDGs multiclass clasiffier," Sep. 2022. [Online]. Available: https://zenodo.org/record/7095783
[30] A. Dubey, A. Jauhri, and A. Pandey, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783
[31] Mistral AI Team, "Mistral small 3.1," https://mistral.ai/news/mistral-small-3-1, Mar. 2025, accessed: 2025-06-12.
[32] Qwen, :, A. Yang, and B. Yang, "Qwen2.5 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.15115
[33] I. Anthropic, "The claude 3 model family: Opus, sonnet, haiku," *Anthropic Research Papers*, 2024, available at https://www.anthropic.com.
[34] J. Ye, X. Chen, and N. Xu, "A comprehensive capability analysis of gpt-3 and gpt-3.5 series models," *arXiv preprint arXiv:2303.10420*, 2023.
[35] OpenAI, J. Achiam, S. Adler, and S. Agarwal, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774
[36] T. Kojima, S. S. Gu, and M. Reid, "Large Language Models are Zero-Shot Reasoners," Jan. 2023, arXiv:2205.11916 [cs]. [Online]. Available: http://arxiv.org/abs/2205.11916
[37] J. Wei, X. Wang, and D. Schuurmans, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903
[38] A. Lacoste, A. Luccioni, and V. Schmidt, "Quantifying the carbon emissions of machine learning," *arXiv preprint arXiv:1910.09700*, 2019.
[39] R. Aiyappa, J. An, and H. Kwak, "Can we trust the evaluation on ChatGPT?" in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, and R. Gupta, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 47–54. [Online]. Available: https://aclanthology.org/2023.trustnlp-1.5
[40] Y. Liu, D. Iter, and Y. Xu, "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment," May 2023, arXiv:2303.16634 [cs]. [Online]. Available: http://arxiv.org/abs/2303.16634
[41] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013
[42] J. Wang, Y. Liang, and F. Meng, "Is ChatGPT a Good NLG Evaluator? A Preliminary Study," Oct. 2023, arXiv:2303.04048 [cs]. [Online]. Available: http://arxiv.org/abs/2303.04048

APPENDIX

A. Prompt Template : Synonym Generation

```
{
  'instructions': [
    'I am asking you to play the role of
   English translator. I am going to give you
   a keyword and I want you to generate 3
   different synonyms. I do not want any
   explanations, formatting or even acronyms
   or abbreviations, just a text that
   everyone can understand.',
    'Synonyms must follow the following
   output format: [synonym1, synonym2,
   synonym3].',
    'I want the synonyms I require and
   nothing else.'
  ],
  'keyword': '<keyword>',
  'output': ''
}
```

B. Prompt Template : Data Reformation

TABLE IV

Prompt	**Instruction**
$formal\_rewrite$	Rephrase the following text using formal, professional language suitable for an academic or institutional context. Ensure the rewritten version highlights themes related to <SDG>: <text>
$simplified\_technical\_summary$	Convert the following text into a simplified explanation, suitable for a general educated audience, while keeping the focus on topics related to <SDG>: <text>
$third\_person\_perspective$	Rewrite the following text from a third-person perspective, avoiding personal references. Emphasize aspects related to <SDG>: <text>
$analytical\_style$	Reformulate the following text in an analytical, data-driven style, ensuring it connects to the theme of <SDG>: <text>
$narrative\_compression$	Transform the following text into a tightly focused, information-dense paragraph that prioritizes key ideas. Make sure it remains centered on the topic of <SDG>: <text>
$descriptive\_expository\_style$	Rewrite the following text in a descriptive, expository style. Focus on presenting clear facts and details related to <SDG>: <text>
$formal\_comparative\_framing$	Reframe the following text to include comparison or contrast where possible. Highlight its relevance to <SDG>: <text>
$objective\_policy\_brief\_style$	Recast the following text in the style of a policy brief using neutral language. Emphasize themes that align with <SDG>: <text>

C. Prompt Template : Training-Label Generation

```
{
    'instructions': [
        'Classify the following text into
    only one of the categories related to the
    Sustainable Development Goals (SDGs).',
        'Use the category definitions below
    to guide your classification.',
        'Only return the SDG category name
    (e.g., SDG_4_QUALITY_EDUCATION), without
    explanations or any additional text.',
        'Category: SDG_1_NO_POVERTY -
    Definition: End poverty in all its forms
    everywhere.',
        'Category: SDG_2_ZERO_HUNGER -
    Definition: End hunger, achieve food
    security and improved nutrition, and
    promote sustainable agriculture.',
        'Category:
    SDG_3_GOOD_HEALTH_AND_WELL_BEING -
    Definition: Ensure healthy lives and
    promote well-being for all at all ages.',
        'Category: SDG_4_QUALITY_EDUCATION -
    Definition: Ensure inclusive and equitable
    quality education and promote lifelong
    learning opportunities for all.',
        'Category: SDG_5_GENDER_EQUALITY -
    Definition: Achieve gender equality and
    empower all women and girls.',
        'Category:
    SDG_6_CLEAN_WATER_AND_SANITATION -
    Definition: Ensure availability and
    sustainable management of water and
    sanitation for all.',
        'Category:
    SDG_7_AFFORDABLE_AND_CLEAN_ENERGY -
    Definition: Ensure access to affordable,
    reliable, sustainable and modern energy
    for all.',
        'Category:
    SDG_8_DECENT_WORK_AND_ECONOMIC_GROWTH -
    Definition: Promote sustained, inclusive
    and sustainable economic growth, full and
    productive employment and decent work for
    all.',
        'Category: SDG_9_INDUSTRY_INNOVATION_
        AND_INFRASTRUCTURE - Definition:
    Build resilient infrastructure, promote
    inclusive and sustainable
    industrialization and foster innovation.',
        'Category:
    SDG_10_REDUCED_INEQUALITIES - Definition:
    Reduce inequality within and among
    countries.',
        'Category:
    SDG_11_SUSTAINABLE_CITIES_AND_COMMUNITIES
    - Definition: Make cities and human
    settlements inclusive, safe, resilient and
    sustainable.',
        'Category:
    SDG_12_RESPONSIBLE_CONSUMPTION_
        AND_PRODUCTION - Definition: Ensure
    sustainable consumption and production
    patterns.',
        'Category: SDG_13_CLIMATE_ACTION -
    Definition: Take urgent action to combat
    climate change and its impacts.',
        'Category: SDG_14_LIFE_BELOW_WATER -
    Definition: Conserve and sustainably use
    the oceans, seas and marine resources for
    sustainable development.',
        'Category: SDG_15_LIFE_ON_LAND -
    Definition: Protect, restore and promote
    sustainable use of terrestrial ecosystems,
    manage forests sustainably, combat
    desertification, and halt biodiversity
    loss.',
        'Category: SDG_16_PEACE_JUSTICE_
        AND_STRONG_INSTITUTIONS - Definition:
    Promote peaceful and inclusive societies,
    provide access to justice for all, and
    build effective, accountable institutions
    at all levels.',
        'Category:
    SDG_17_PARTNERSHIPS_FOR_THE_GOALS -
    Definition: Strengthen the means of
    implementation and revitalize the global
    partnership for sustainable development.'
     ],
     'text': '<text>',
     'classification': ''
}
```

D. BERT Training Parameters

TABLE V

Parameters	Values
batch_size	16
lr	2e-5
max_seq_length	512
optimizer	adamw_torch
weight_decay	0.01
scheduler	cosine
warmup_ratio	0.1
epochs	3
mixed_precision	bf16
eval_strategy	epoch
auto_find_batch_size	false
gradient_accumulation	1
logging_steps	-1
max_grad_norm	1
save_total_limit	1
seed	42

TABLE VI: Per-class analysis of method performance across SDGs (precision / recall)

(a) SDG 1-4

Type	Method	SDG 1	SDG 2	SDG 3	SDG 4
SDG Specific	SDG Mapper	0.63 / 0.77	0.82 / 0.68	0.88 / 0.65	0.79 / 0.45
	Aurora	0.41 / 0.86	0.65 / 0.77	0.83 / 0.35	0.91 / 0.47
	Elsevier	0.30 / 0.94	0.71 / 0.63	0.74 / 0.17	0.86 / 0.82
ZS-NLI	BART-Large-MNLI	0.41 / 0.58	0.74 / 0.32	0.66 / 0.76	0.89 / 0.48
Open LLMs	Llama 3.2	0.84 / 0.63	0.96 / 0.68	0.88 / 0.76	0.76 / 0.89
	Llama 3.1	0.73 / 0.93	0.98 / 0.73	0.93 / 0.91	0.69 / 0.95
	Mistral Small 3.1	0.80 / 0.93	0.91 / 0.95	0.99 / 0.95	0.86 / 0.93
	Gemma 3	0.84 / 0.93	0.94 / 0.93	0.99 / 0.96	0.91 / 0.97
	Qwen2.5	0.79 / 0.94	0.90 / 0.91	0.96 / 0.96	0.90 / 0.96
Closed LLMs	Claude 3.5 Haiku	0.78 / 0.98	0.96 / 0.81	0.96 / 0.96	0.91 / 0.97
	Mistral Large 2	0.84 / 0.93	0.93 / 0.93	0.95 / 0.97	0.90 / 0.95
	GPT-3.5-Turbo	0.84 / 0.93	0.86 / 0.91	0.95 / 0.97	0.91 / 0.95
	GPT-4o	0.87 / 0.93	0.95 / 0.93	0.99 / 0.95	0.92 / 0.96
Our	LëtzSDG	0.73 / 0.92	0.86 / 0.94	0.95 / 0.91	0.89 / 0.95

(b) SDG 5-8

Type	Method	SDG 5	SDG 6	SDG 7	SDG 8
SDG Specific	SDG Mapper	0.96 / 0.44	0.84 / 0.75	0.80 / 0.67	0.42 / 0.65
	Aurora	0.78 / 0.78	0.77 / 0.74	0.87 / 0.63	0.34 / 0.80
	Elsevier	0.78 / 0.76	0.77 / 0.75	0.94 / 0.34	0.22 / 0.54
ZS-NLI	BART-Large-MNLI	0.91 / 0.66	0.91 / 0.59	0.63 / 0.44	0.27 / 0.74
Open LLMs	Llama 3.2	0.95 / 0.78	0.64 / 0.82	0.91 / 0.70	0.74 / 0.53
	Llama 3.1	0.94 / 0.89	0.85 / 0.88	0.91 / 0.79	0.41 / 0.82
	Mistral Small 3.1	0.95 / 0.89	0.91 / 0.91	0.94 / 0.94	0.78 / 0.78
	Gemma 3	0.95 / 0.90	0.90 / 0.90	0.91 / 0.87	0.54 / 0.84
	Qwen2.5	0.94 / 0.91	0.92 / 0.90	0.91 / 0.93	0.56 / 0.80
Closed LLMs	Claude 3.5 Haiku	0.95 / 0.90	0.94 / 0.90	0.94 / 0.90	0.59 / 0.81
	Mistral Large 2	0.96 / 0.89	0.90 / 0.90	0.94 / 0.87	0.65 / 0.80
	GPT-3.5-Turbo	0.95 / 0.85	0.84 / 0.97	0.94 / 0.90	0.61 / 0.88
	GPT-4o	0.96 / 0.93	0.94 / 0.91	0.96 / 0.93	0.77 / 0.84
Our	LëtzSDG	0.94 / 0.90	0.88 / 0.93	0.92 / 0.90	0.64 / 0.80

(c) SDG 9-12

Type	Method	SDG 9	SDG 10	SDG 11	SDG 12
SDG Specific	SDG Mapper	0.87 / 0.41	0.62 / 0.37	0.78 / 0.51	0.88 / 0.70
	Aurora	0.60 / 0.55	0.43 / 0.25	0.62 / 0.56	0.50 / 0.50
	Elsevier	0.78 / 0.40	0.23 / 0.16	0.57 / 0.55	0.59 / 0.64
ZS-NLI	BART-Large-MNLI	0.65 / 0.46	0.50 / 0.39	0.50 / 0.46	0.44 / 0.65
Open LLMs	Llama 3.2	0.54 / 0.70	0.61 / 0.75	0.75 / 0.50	0.47 / 0.82
	Llama 3.1	0.82 / 0.52	0.95 / 0.63	0.82 / 0.74	0.88 / 0.83
	Mistral Small 3.1	0.92 / 0.78	0.85 / 0.78	0.77 / 0.87	0.91 / 0.83
	Gemma 3	0.84 / 0.74	0.87 / 0.68	0.86 / 0.84	0.94 / 0.86
	Qwen2.5	0.92 / 0.70	0.90 / 0.75	0.86 / 0.81	0.94 / 0.83
Closed LLMs	Claude 3.5 Haiku	0.86 / 0.79	0.92 / 0.70	0.87 / 0.82	0.93 / 0.88
	Mistral Large 2	0.86 / 0.82	0.85 / 0.75	0.77 / 0.93	0.93 / 0.78
	GPT-3.5-Turbo	0.83 / 0.85	0.91 / 0.71	0.83 / 0.81	0.84 / 0.88
	GPT-4o	0.89 / 0.82	0.87 / 0.81	0.85 / 0.94	0.87 / 0.89
Our	LëtzSDG	0.82 / 0.74	0.85 / 0.66	0.76 / 0.75	0.91 / 0.84

(d) SDG 13-17

Type	Method	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17
SDG Specific	SDG Mapper	0.88 / 0.75	0.98 / 0.63	0.94 / 0.77	0.86 / 0.67	0.82 / 0.59
	Aurora	0.76 / 0.68	0.78 / 0.40	0.74 / 0.64	0.71 / 0.52	0.23 / 0.74
	Elsevier	0.46 / 0.49	0.69 / 0.90	0.93 / 0.64	0.52 / 0.77	0.00 / 0.00
ZS-NLI	BART-Large-MNLI	0.60 / 0.84	0.59 / 0.41	0.42 / 0.56	0.68 / 0.55	0.59 / 0.76
Open LLMs	Llama 3.2	0.47 / 0.89	1.00 / 0.67	1.00 / 0.43	0.71 / 0.73	0.88 / 0.81
	Llama 3.1	0.75 / 0.83	1.00 / 0.78	0.94 / 0.77	0.80 / 0.85	0.81 / 0.89
	Mistral Small 3.1	0.88 / 0.92	1.00 / 0.90	0.95 / 0.95	0.88 / 0.92	0.89 / 0.91
	Gemma 3	0.87 / 0.90	1.00 / 0.90	0.95 / 0.92	0.88 / 0.92	0.76 / 0.94
	Qwen2.5	0.86 / 0.84	1.00 / 0.84	0.92 / 0.95	0.88 / 0.89	0.72 / 0.94
Closed LLMs	Claude 3.5 Haiku	0.88 / 0.91	1.00 / 0.92	0.95 / 0.93	0.87 / 0.95	0.81 / 0.94
	Mistral Large 2	0.87 / 0.93	0.99 / 0.90	0.95 / 0.92	0.92 / 0.87	0.89 / 0.91
	GPT-3.5-Turbo	0.89 / 0.88	0.97 / 0.92	0.99 / 0.85	0.94 / 0.89	0.86 / 0.91
	GPT-4o	0.88 / 0.92	1.00 / 0.92	0.97 / 0.92	0.88 / 0.93	0.84 / 0.94
Our	LëtzSDG	0.79 / 0.85	0.99 / 0.90	0.97 / 0.92	0.92 / 0.80	0.66 / 0.80