Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

Classifying and Tracking International Aid Contribution Towards SDGs

Sungwon Park[∗1,2], **Dongjoon Lee**[∗1], **Kyeongjin Ahn**[1,2], **Yubin Choi**[1],
Junho Lee[3], **Meeyoung Cha**[2,1] and **Kyung Ryul Park**[1]

[1] Korea Advanced Institute of Science and Technology (KAIST)
[2] Max Planck Institute for Security and Privacy (MPI-SP)
[3] United Nations (UN)

Abstract

International aid is a critical mechanism for promoting economic growth and well-being in developing nations, supporting progress toward the Sustainable Development Goals (SDGs). However, tracking aid contributions remains challenging due to labor-intensive data management, incomplete records, and the heterogeneous nature of aid data. Recognizing the urgency of this challenge, we partnered with government agencies to develop an AI model that complements manual classification and mitigates human bias in subjective interpretation. By integrating SDG-specific semantics and leveraging prior knowledge from language models, our approach enhances classification accuracy and accommodates the diversity of aid projects. When applied to a comprehensive dataset spanning multiple years, our model can reveal hidden trends in the temporal evolution of international development cooperation. Expert interviews further suggest how these insights can empower policymakers with data-driven decision-making tools, ultimately improving aid effectiveness and supporting progress toward SDGs.

1 Introduction

International aid, commonly known as Official Development Assistance (ODA), refers to financial support provided by governments to help developing countries combat poverty and promote economic growth. For decades, international aid has played a crucial role in socioeconomic development, particularly in advancing the United Nations Sustainable Development Goals (SDGs), by providing essential resources to vulnerable populations [Alsayyad, 2020; Hynes and Scott, 2013]. Ensuring the effectiveness of aid and progress toward the SDGs requires rigorous tracking of funding allocation and impact assessment.

The Creditor Reporting System (CRS), established by the Organization for Economic Cooperation and Development (OECD) in 1967, is the most authoritative and comprehensive database that records decades of data on international aid projects. This database provides consistent and reliable statistics on development finance and individual aid projects [UNAIDS, 2004; Morgenstern and Brown, 2022]. The CRS is publicly accessible, ensuring transparency for policymakers, researchers, and civil society organizations (CSOs). Each year, OECD member countries are required to report project-level information on their aid activities, which includes 93 data features such as *project description*, *donor code*, and *recipient code*, *purpose code* (sector), and *commitment value* (budget). CRS receives over 250,000 new project entries annually, detailing activities in various recipient countries. Since 2018, the system has asked donors to specify the relevant SDG (*SDG focus index*) for each project to assess its alignment with global goals [OECD, 2018; Kasneci, 2018]. However, this manual classification has lacked consistency and systematic rigor, owing to its labor-intensive, costly, and inherently subjective nature. Moreover, the absence of SDG-related records prior to 2018, coupled with the fact that over half of reported projects since then remain uncategorized, presents significant challenges to assessing aid effectiveness. These limitations hinder a comprehensive understanding of how aid impacts the SDGs.

Major development agencies, including the United Nations (UN) and OECD, are exploring methods to automate the classification of aid projects and link them to the SDGs [Ericsson and Mealy, 2019; Pincet *et al.*, 2019; Lee *et al.*, 2023]. However, these efforts are complex due to the heterogeneous nature of the aid projects contributed by various countries and agencies. This diversity leads to significant syntactic and semantic variations in the project descriptions in CRS. Subjective judgments and sociopolitical contexts further influence how projects are categorized under specific aid goal, often resulting in misalignment between project descriptions and their assigned categories. These inconsistencies hinder classifiers from generalizing to unseen projects, particularly those related to underrepresented SDGs. We validated these rationales through interviews with aid experts.

As part of a joint effort between government experts and AI researchers to automate aid categorization, we identified key demands and developed a multi-label text classifier to improve the mapping of SDG impact. By integrating their semantic definitions into the classifier's encoder, our goal was to support aid management and targeting while minimizing the influence of human bias in subjective interpretation of aid projects. Our classifier learns data semantics through self-supervised learning and leverages insights from the large language model (LLM) to incorporate pre-trained knowledge. The training objective is to improve generalization, particularly for underrepresented SDGs, by classifying previously unseen project description sentences during training.[1]

We validated the AI classification results with aid experts and confirmed that our approach demonstrates exceptional performance in identifying relevant SDG categories. It effectively captures the complexities of real-world projects by filling in the past and missing data on a comprehensive scale. This capability also allows the analysis of aid financing trends from a longitudinal perspective, providing valuable information to policymakers. We present our findings from expert interviews on AI classification and its practical usefulness. Building on this achievement, we emphasize the need for a global effort to develop objective, transparent, and universally accepted standards for aid project classification through collaboration between aid experts and AI researchers.

2 Related Work

2.1 AI in the SDG Monitoring

Tracking the progress of aid projects is essential for organizations such as the UN and the OECD. However, organizing development aid projects to align with the appropriate goals presents several challenges. These include: 1) the complexity of multi-label classification, which necessitates a comprehensive understanding of each project across 93 features and reasoning to map them to multiple SDG goals based on the 169 official targets; 2) interdependence and trade-offs between 17 SDGs; and 3) the imbalanced distribution of goals and data classes [Kasneci, 2018]. Previous research applied machine learning to categorize purpose codes within development aid projects [Ericsson and Mealy, 2019; Lee *et al.*, 2023; Pincet *et al.*, 2019], leading to reduced expert labor and improved precision [Lee *et al.*, 2023]. However, previous approaches have been limited to single-label classification or have failed to account for the complexities of multi-label classification and semantic information, such as country-specific descriptions and SDG definitions, which are essential for accurately representing development aid projects.

2.2 Multi-Label Classification

In multi-label text classification, several models have been proposed to capture label dependencies, including classifier chains [Read *et al.*, 2009], graph neural networks (GNNs) [Zhang and Chen, 2018], and label-specific attention networks [Xiao *et al.*, 2019]. Among them, classifier chains are susceptible to error propagation, where a misclassification in one label affects subsequent predictions. Alternatively, transformer-based models that utilize contextualized semantic understanding from pre-training on large corpora have been applied to improve multi-label classification. For instance, BERT has been used for document classification [Devlin *et al.*, 2018; Lee *et al.*, 2021]. These models enhance classification robustness by leveraging the pre-trained knowledge and handling complex textual data more effectively.

[1]Our code is made freely available to support broader use at https://github.com/deu30303/ODA_SDG

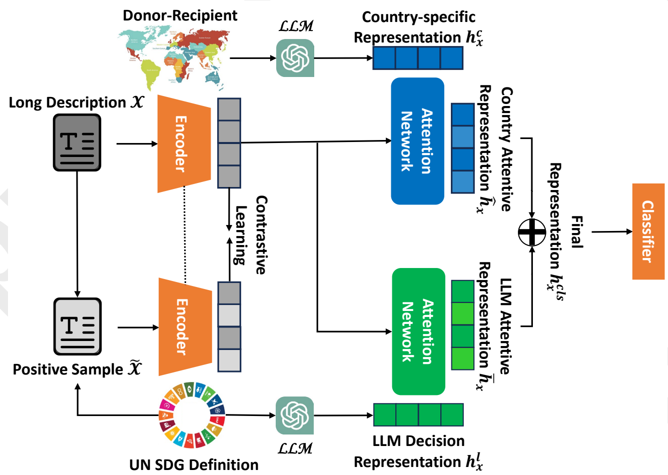

Figure 1: The proposed framework has three main components: (1) the SDG Semantics Injection Module; (2) the Country Information Guided Module; and (3) the LLM Decision Guided Module.

However, while BERT excels at capturing the bidirectional context of text, it does not explicitly model label correlations. To address this limitation, recent methods have integrated label co-occurrence into text encoders, improving their ability to capture inter-label dependencies and enhancing multi-label classification performance. GCLR employs contrastive learning to incorporate label graph structures into semantic representations [Wang *et al.*, 2022]. In this study, we leverage LLM prior knowledge instead of label graph information to provide a more generalizable understanding of label relationships, thereby enhancing multi-label classification tasks.

3 Method

Problem Statement: We tackle a multi-label classification problem to determine the most relevant goals from the list of 17 SDGs for each aid project. This classification relies on project descriptions, meta-information such as donor and recipient countries, and the needs of different sectors and policies. Given a project description x in natural language, our objective is to predict the corresponding multi-label y.

Overview. Figure 1 presents an overall framework, which aims to enhance the final representation $\mathbf{h}_x^{cls}$ derived from x by integrating the semantic information of the SDGs into task-specific classifiers, such as BERT. The framework comprises three core modules, detailed below:

- **Semantics Injection Module** (§**3.1**)**:** Building on expert manual labeling guided by SDG definitions, we integrate the semantic representations $\mathbf{h}_g$ of the goals into the token embedding set $\mathcal{E}$ from the BERT encoder through contrastive learning. We sample essential tokens to determine aid goals and use them to construct positive samples $\tilde{x}$ in contrastive learning.
- **Country Information Guided Module** (§**3.2**)**:** Given the key information on development aid policy for donor and recipient countries, we derive an attentive representation

$\hat{\mathbf{h}}_x$ that explicitly incorporates these policies to account for political contexts shaped by donor-recipient relationships.

- **LLM Decision Guided Module (§3.3):** When the classifier encounters sentences it has not seen during training, its generalization ability, particularly for underrepresented SDGs, may be limited. We learn an attentive representation $\bar{\mathbf{h}}_x$ derived from the LLM decision. We predict the usefulness of the LLM decision α_x in the final representation (i.e., $h_x^{cls} = \hat{\mathbf{h}}_x + \alpha_x \bar{\mathbf{h}}_x$).

3.1 Semantics Injection Module

During the initial phase of aid decisions, guidelines are typically established to define target objectives that align with sectoral priorities and policy needs. Aid experts then refer to these specific definitions and descriptions to determine the relevant SDG focus. However, this is nontrivial as projects are often framed in broad terms that emphasize overarching goals, such as eradicating extreme poverty (SDG Goal 1).

As a result, when a text classifier is trained solely with traditional supervised loss functions (e.g., binary cross-entropy), it fails to incorporate the semantic meaning of the SDGs. Our objective is to develop SDG-aware text representations using contrastive learning to identify latent patterns related to each SDG's semantic definition, while also extracting discriminative features that distinguish between the goals [Park *et al.*, 2022; Gao *et al.*, 2021]. We use goal-semantic representations to guide the construction of positive samples $\tilde{x}$ for contrastive learning.

Goal 1: End poverty in all its forms everywhere
Target 1.a: Ensure significant mobilization of resources from a variety of sources, ...

Goal 2: End hunger, achieve food security and improved nutrition, and promote sustainable agriculture
Target 2.a: Increase investment, including through enhanced international cooperation, ...

Table 1: Example SDG definition and its corresponding target.

Positive Sample Construction. Given project description x, let $\mathcal{E}$ denote the token embedding set from BERT as follows:

$$\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\} \quad \text{(Eq. 1)}$$

where $\mathbf{e}_i$ is the i-th token embedded in the BERT encoder. We aim to construct the positive sample for $\tilde{x}$ by sampling salient tokens conditioned on the semantic representations of the goal definition. These semantic representations are derived from the official definitions of each SDG and their corresponding targets (see Table 1 for an example). The representation of the j-th goal, $\mathbf{h}_{g_j}$, can also be extracted using BERT embedding. The scale-dot attention weight A is utilized to measure the relevance between token embedding $\mathbf{e}_i$ and goal embedding $\mathbf{g}_j$:

$$Q_i = W_Q \mathbf{e}_i, \; K_j = W_K \mathbf{h}_{g_j}, \; A_{ij} = \frac{Q_i K_j^T}{\sqrt{d_h}} \quad \text{(Eq. 2)}$$

where W_Q and W_K are two weight matrices used for the attention mechanism. For simplicity, subscripts of all independent attention matrices are omitted.

We sample salient tokens conditioned on the multi-label y by computing the probability distribution P with Gumbel-Softmax for differentiable sampling [Jang *et al.*, 2017] (Eq. 4). Here, each token can contribute to more than one goal, so the importance score P_i for the i-th token, with respect to the multilabel y is computed by summing the probabilities of all ground-truth labels present in y as follows:

$$P_{ij} = Gumbel(A_{i1}, A_{i2}, , ..., A_{i17})_j \quad \text{(Eq. 3)}$$

$$P_i = \sum_{j \in y} P_{ij} \quad \text{(Eq. 4)}$$

Finally, we construct a positive sample $\tilde{x}$ by retaining only the salient tokens with thresold τ and replacing all other irrelevant tokens with 0:

$$\tilde{x}_i = \begin{cases} x_i & \text{if } P_i > \tau \\ 0 & \text{else} \end{cases} \quad \text{(Eq. 5)}$$

Semantics-Aware Contrastive Learning. The next objective is to train a text encoder that incorporates goal semantic information into its representations, using contrastive learning with the given positive sample $\tilde{x}$. Let h_x denote the average pooling of token representation of x fed by BERT (i.e., $\mathbf{h}_x = \text{AvgPool}(\mathcal{E})$). Even after removing irrelevant tokens, the positive sample constructed from goal representations retains a semantic context similar to the original sentence. A shallow MLP layer, S_G, is added on top of the representation space to ensure that the sentences embedded in similar sentences are brought closer together, while the negative sample embedding h_{x^n} from different sentences in the batch $\mathcal{X}$ are pushed apart through contrastive learning:

$$\mathcal{L}_G = -\log \frac{\exp(\text{sim}(S_G(\mathbf{h}_x), S_G(\mathbf{h}_{\tilde{x}})))}{\sum_{x^n \in \mathcal{X}} \exp(\text{sim}(S_G(\mathbf{h}_x), S_G(\mathbf{h}_{x^n})))} \quad \text{(Eq. 6)}$$

where sim represents the cosine similarity.

3.2 Country Information Guided Module

Understanding the political contexts between donor and recipient countries is essential, as they expose the underlying power structures and biases that shape interdependence among multi-labels. For example, donor's aid allocation often reflects commercial or geopolitical objectives. Similarly, each country's SDG strategies are influenced by various institutional challenges and require a contextualized approach [Park and Park, 2024]. Thus, relying solely on semantic information from project descriptions during training is insufficient. To account for the complex doner-recipient relationships, we first extract insights derived from LLM on the countries and integrate this information using an attention mechanism to guide the text classifier.

Extracting Insights on Countries. We design an instruction prompt to generate responses that explore the key aid policies of each donor and the development strategies of the recipient country, as in Table 2. We include the income level

Based on government documents, summarize the Official Development Assistance (ODA) Policy of [donor country]?

In the OECD CRS (Creditor Reporting System) database, income groups categorize countries based on their gross national income (GNI) per capita. Here's what each group represents: - LICs (Low-Income Countries): Countries with a GNI per capita of $1,045 or less. - LMICs (Lower Middle-Income Countries): Countries with a GNI per capita between $1,046 and $4,095. - UMICs (Upper Middle-Income Countries): Countries with a GNI per capita between $4,096 and $12,695. - LDCs (Least Developed Countries): This group includes countries identified by the United Nations as facing severe structural impediments to sustainable development. Based on credible government papers, can you summarize the Official Development Assistance (ODA) Policy of [income classification], [recipient country]?

Table 2: Prompt example used in research. The "[country]" part will be replaced by the country name, and "[income classification]" will indicate the recipient country's GNI per capita level.

of the recipient country in prompts because of a significant connection between the two variables [Knack *et al.*, 2011; Easterly, 2014]. The income group, which best represents aid demand and absorptive capacity of a recipient country, serves as a primary criterion for effective allocation of aid [Collier and Dollar, 2002; Dalgaard *et al.*, 2004; Easterly and Pfutze, 2008]. For example, for lower-income countries, the focus of aid tends to be on poverty alleviation, basic needs, and infrastructure. Middle-income countries may prioritize advanced developmental outcomes such as institutional development, technological innovation, and economic governance for growth.

Country Information-Driven Attention. Given text input x and its corresponding model responses regarding donor and recipient countries, we extract the donor representation $\mathbf{h}_x^d$ and the recipient representation $\mathbf{h}_x^r$ from BERT encoder. The country-specific representation $\mathbf{h}_x^c$ is then constructed by combining the donor and recipient representations through element-wise multiplication (see Eq. 7). To incorporate this country information into the text representation, we introduce cross-attention between the token embedding set $\mathcal{E}$ and the country representation $\mathbf{h}_x^c$ as follows:

$$\mathbf{h}_x^c = S_C(\mathbf{h}_x^d \odot \mathbf{h}_x^r) \quad \text{(Eq. 7)}$$

$$Q_i = W_Q \mathbf{e}_i,\ K = W_K \mathbf{h}_x^c,\ A_i = \frac{Q_i K^T}{\sqrt{d_h}} \quad \text{(Eq. 8)}$$

$$\hat{A} = \text{Softmax}(A) \quad \text{(Eq. 9)}$$

$$\hat{\mathbf{h}}_x = \text{AvgPool}(\{\hat{A}_1\mathbf{e}_1, \hat{A}_2\mathbf{e}_2, ..., \hat{A}_n\mathbf{e}_n\}) \quad \text{(Eq. 10)}$$

where S_C is the MLP layer and AvgPool is the average pooling applied to the token representations.

We append an auxiliary classifier f_c to the attentive representation $\hat{\mathbf{h}}_x$ with the conventional binary cross-entropy loss H to directly perform multi-label classification. The refined representation $\hat{\mathbf{h}}_x$ plays an important role in deciding on which tokens to focus on based on the country information.

$$\mathcal{L}_C = H(y, f_c(\hat{\mathbf{h}}_x)) \quad \text{(Eq. 11)}$$

Now you will help me classify the sentence according to the following 17 Sustainable Development Goals (SDGs): **Goal 1:** End poverty in all its forms everywhere ... < EXAMPLES >: [Project description] Determine which of the above Sustainable Development Goals (SDGs) does the < EXAMPLES > correspond to. You can select multiple answers following format. < Answer >: < 1 ∼ 17 >

Table 3: Prompt to extract the language model decision for each project description: The "[Project description]" part will be filled with the specific project description.

3.3 LLM Decision Guided Module

The CRS is the official databse that incorporates input from a wide range of countries and experts on international aid. However, this inclusivity leads to varied syntactic and semantic distributions in the textual data, presenting significant challenges for multilabel classifiers. Limited training data for underrepresented SDGs further limit the generalizability of task-specific classifiers. In contrast, LLMs are known to demonstrate broader generalization capabilities due to their higher representational capacity, which we integrate using an attention mechanism.

Decision-Driven Attention. We first extract the LLM decision $\bar{y}$ for the project description x by building a prompt using the definition of each goal, as illustrated in Table 3. The decision representation $\mathbf{h}_x^l$ is constructed by sampling the goal representation corresponding to the decision $\bar{y}$ combining them through multiplication, and passing the result through an MLP layer S_L (Eq. 12). Similarly to incorporating country information, we applied a mechanism of cross-attention between the token embedding set and the decision representation as follows:

$$\mathbf{h}_x^l = S_L(\prod_{j \in \bar{y}} \mathbf{h}_{g_j}) \quad \text{(Eq. 12)}$$

$$Q_i = W_Q \mathbf{e}_i,\ K = W_K \mathbf{h}_x^l,\ A_i = \frac{Q_i K^T}{\sqrt{d_h}} \quad \text{(Eq. 13)}$$

$$\bar{A} = \text{Softmax}(A) \quad \text{(Eq. 14)}$$

$$\bar{\mathbf{h}}_x = \text{AvgPool}(\{\bar{A}_1\mathbf{e}_1, \bar{A}_2\mathbf{e}_2, ..., \bar{A}_n\mathbf{e}_n,\}) \quad \text{(Eq. 15)}$$

Consistent with Eq. 11, we apply standard binary cross-entropy loss using the auxiliary classifier f_l on top of the attentive representation $\bar{h}_x$:

$$\mathcal{L}_D = H(y, f_l(\bar{\mathbf{h}}_x)) \quad \text{(Eq. 16)}$$

Decision Evaluation. Naive utilization of LLM decision can result in suboptimal prediction performance. To address this issue, we propose an additional evaluation process that assesses the usefulness of decision and calibrates their weights for subsequent multi-label predictions. We generate the usefulness label y_u by determining whether the original label y overlaps with the decision $\bar{y}$. We evaluated the effectiveness of the decision-driven representation of the model $\bar{\mathbf{h}}_x$ using this usefulness label y_u. We train the usefulness classifier f_u with binary cross-entropy loss:

$$\mathcal{L}_U = H(y_u, f_u(\bar{\mathbf{h}}_x)) \quad \text{(Eq. 17)}$$

Multi-label Prediction with Calibrated Representation. Using the usefulness classifier f_u , we determine the scaling coefficient α_x by evaluating the usefulness of the LLM decision-driven representation $\bar{\mathbf{h}}_x$ based on the prediction probability of the classifier (that is, $\alpha_x = f_u(\bar{\mathbf{h}}_x)$). The refined final representation h_x^{cls} is derived by combining the country information-driven representation $\hat{h}_x$ with the calibrated decision-driven representation $\bar{\mathbf{h}}_x$ (Eq. 18). Then the final representation $\mathbf{h}_x^{cls}$ is then passed through the MLP classifier f_{cls} for final classification prediction:

$$h_x^{cls} = \hat{\mathbf{h}}_x + \alpha_x \bar{\mathbf{h}}_x \quad \text{(Eq. 18)}$$

$$\mathcal{L}_{CE} = H(y, f_{cls}(\mathbf{h}_x^{cls})) \quad \text{(Eq. 19)}$$

Finally, the overall loss function is computed by combining the loss terms discussed above with hyperparameters λ_1, λ_2 as given in Eq. 20.

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_G + \lambda_2(\mathcal{L}_C + \mathcal{L}_D + \mathcal{L}_U) \quad \text{(Eq. 20)}$$

For inference, we extract country information and decision for the test data x_t. We then perform an element-wise summation of the refined characteristics: the country information-driven representation $\hat{\mathbf{h}}_{x_t}$ and the calibrated decision-driven representation $\alpha_{x_t}\bar{\mathbf{h}}_{x_t}$ to obtain the final output characteristic (i.e., $h_{x_t}^{cls} = \hat{\mathbf{h}}_{x_t} + \alpha_{x_t}\bar{\mathbf{h}}_{x_t}$). The final feature is then passed to the classifier f_{cls}.

4 Experiment

4.1 Performance Evaluation

Dataset. As introduced in Section 1, our main dataset, the CRS is internationally recognized as the standard for detailed project-level aid data, providing breakdowns by country, sector, type of aid, donor, and year. Based on consultations with aid experts, among the 93 data fields, we selected four key features: *project description*, *SDG focus index*, *donor code*, and *recipient code*. The *project description* provides a detailed text of up to 4,000 words, used to classify the project sector and its alignment with the aid goal. Introduced in 2018, *SDG focus index* is a multi-entry label indicating up to ten SDGs the project focuses on. However, reporting remains voluntary due to the significant manpower required for data entry and review [OECD, 2018].

The labels y correspond to the 17 SDGs listed in this focus index. Figure 2 shows the data statistics, indicating under-represented goals such as SDG 7 and SDG 14. The *donor code* and *recipient code* are the nations participating in each aid project. The former is a numerical value, often one of the 32 member countries of the OECD Development Assistance Committee (DAC) or a multilateral donor organization. The latter assigns a numerical value to the 170 recipient countries or territories that receive international aid.

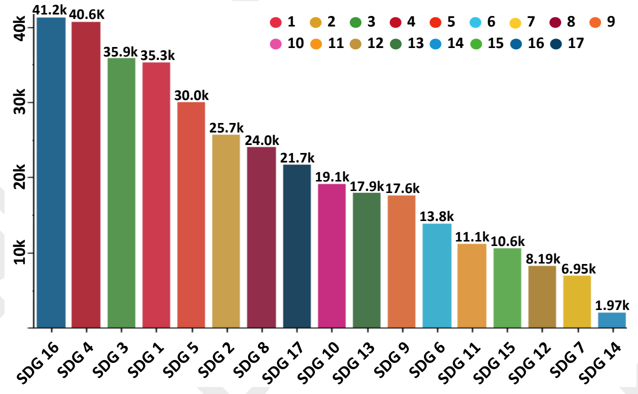

Figure 2: Distribution of SDG focus index by frequency (y).

Implementation. We gathered 151,832 official aid documents and divided them into training and testing sets in a 3:1 ratio. This division was independent of the submission year, ensuring consistency across the data's studied fields. Due to the bilingual nature of the documents (some in English, others in French), we utilized a multilingual BERT as the foundation and trained LLMs over 100 epochs. The Adam optimizer for BERT was configured with a learning rate of 1e-05. The hyperparameters in Eq. 20 were set to 0.1, with a token sampling threshold τ of 0.01. Text-based baselines were evaluated using the same backbone network, optimizer, and learning rate. For graph-based baselines, the adjacency matrix of the label graph was derived from the co-occurrence matrix, which was constructed by counting the frequency of label pairs in the training dataset. This label information was then incorporated into the text encoder. We used OpenAI's GPT-4o-mini with default settings of the top-p value at one and the temperature at zero.

Result. We evaluated our framework against two LLM-based classifiers and three text encoder-based classifiers.

- GPT-4o utilizes in-context learning for inference.
- GPT-4 Turbo uses in-context learning for inference.
- Vanilla BCE is a text encoder-based baseline that applies conventional binary cross-entropy for multi-label classification.
- LGCN integrates label graph information into the text encoder using an attention [Ma *et al.*, 2021].
- GCLR utilizes contrastive learning to integrate label graph information into the text encoder [Wang *et al.*, 2022].

Table 4 reports the performance of multi-label classifiers, showing that our framework consistently surpasses baseline models across all metrics. The table also confirms that text encoder-based classifiers generally perform better than LLM-based classifiers, suggesting the importance of having contextual information.

Model	Precision	Recall	F1-score	AUROC
GPT-4o	0.7409	0.7578	0.7170	-
GPT-4 Turbo	0.7407	0.8049	0.7381	-
Vallia BCE	0.8509	0.8004	0.8243	0.9382
LGCN	0.8628	0.8075	0.8365	0.9521
GCLR	0.8572	0.8142	0.8357	0.9519
Ours	**0.8684**	**0.8192**	**0.8430**	**0.9617**

Table 4: Performance comparison of multi-label classification methods on the CRS dataset. Best result is shown in bold text.

Model	Precision	Recall	F1-score	AUROC
Full Component	0.8684	**0.8192**	**0.8430**	**0.9617**
w/o SDG Semantic	0.8631	0.8121	0.8368	0.9588
w/o Country Info	0.8671	0.8143	0.8399	0.9594
w/o LLM Decision	**0.8787**	0.7971	0.8359	0.9609

Table 5: Ablation study results. The removal of any model component results in a reduction of F1 score and AUROC.

Ablation Study. We conducted ablation experiments in which each component was removed from the entire model and considered four configurations:

- Full Components: Our complete method.
- w/o SDG Semantic: The method excluding contrastive learning loss L_G, as detailed in Eq. 6.
- w/o Country Info : The approach without employing the country information-based attention in Eq. 7 - Eq. 11.
- w/o LLM Decision: The method omitting the language model-specific attention, referenced in Eq. 12 - Eq. 18.

Table 5 shows the performance for each ablation. Each component positively contributes to the improvement of classification performance. The model without the LLM decision achieves higher precision but lower recall, indicating that the SDG Semantics Injection Module and the Country Information Guided Module play a crucial role in enhancing precision, while incorporating the language model's decision further boosts recall. In particular, the LLM decision improved recall for underrepresented classes such as Goal 7 ($0.74 \rightarrow 0.81$) and Goal 14 ($0.68 \rightarrow 0.73$). Due to space limitations, we show results on additional analyses such as the LLM knowledge qualitative analysis and sensitivity analysis for alternative LLM backbones in the Appendix.

5 Implication

We proposed a practical method to map aid projects with the SDGs. By capturing a comprehensive range of project dimensions, our model could improve classification accuracy. Our method is recognized as a substantial contribution to the fields of international development, public policy, and financial management, allowing a more precise evaluation of global development efforts. Moreover, it supports data-driven decision making (DDDM), allowing policy makers to systematically prioritize focus areas and formulate evidence-based policies. We discuss implications for tracking and monitoring aid projects and how AI methods can be used in the strategic planning and execution of development projects.

5.1 Discussion

For financial analysis, we imputed all data from 2015 to 2022 for missing historical aid projects using our model.

Limitations in Analyzing Past Data. As aid categorization has been voluntarily recorded since 2018, there exist no labels prior to that year, and a significant portion of the SDG category data remain missing even after 2018. Much of the data after 2018 suffer from incomplete labels, leading to analyses that may not accurately reflect the actual aid allocations and impacts (see Figure 6 in the Appendix). This issue limits the historical analysis of contributions.

Analysis of Aid Financing Using the Trained Model. To estimate labels for each past and missing project, we first extracted donor-recipient information and decisions through an language model, then applied this information, along with project descriptions, to our trained model. Although a single project may align with multiple goals, distributing the budget equally across all SDGs is suboptimal. We assume that the annual budget sum S_i for the i-th year in the labeled dataset can be expressed as a linear sum of the statistics of the i-th year for each k-th goal, c_i^k, within the labeled dataset. To estimate the budget w_k for each k-th goal, we fit a linear regression model by solving the optimization problem (Eq. 21). The estimated budget proportions, detailed in the Appendix Table 6, are then used in the financing analysis.

$$\min_{w_1,\dots,w_{17}} \sum_{i\in\{2018,..,2022\}} ||S_i - \sum_{k=1}^{17}(c_k^i \times w_k)||_2^2 \quad \text{(Eq. 21)}$$

Our analysis highlights distinct trends from both the recipient and donor perspectives. From the recipient side, illustrated in Figure 3, the allocation of the aid varies by income group. Low-income countries receive concentrated support for foundational goals like SDG 1 (No Poverty) and SDG 2 (Zero Hunger), while lower and upper middle-income countries see increased allocations for goals like SDG 8 (Decent Work and Economic Growth) and SDG 9 (Industry, Innovation, and Infrastructure). This indicates that development needs and economic status shape the distribution of aid, emphasizing the importance of aligning aid with the priorities of different income groups.

Figure 4 shows how donor priorities have shifted across goals, particularly in response to global events over the course of seven years. For example, funding for SDG 3 (Good Health and Well-being) has nearly doubled since 2018 due to the COVID-19 pandemic, before returning to pre-pandemic levels by 2022. This fluctuation demonstrates donors' capacity to adjust priorities in response to urgent global needs. Methods like ours help better understand international development financing, offering insights that were previously unattainable due to extensive missing data and the lack of structured SDG-specific data in aid records.

5.2 Expert Interview

We conducted three semi-structured interviews and two focus group discussions to identify stakeholder's challenges and assess our AI model's relevance and effectiveness. A total of

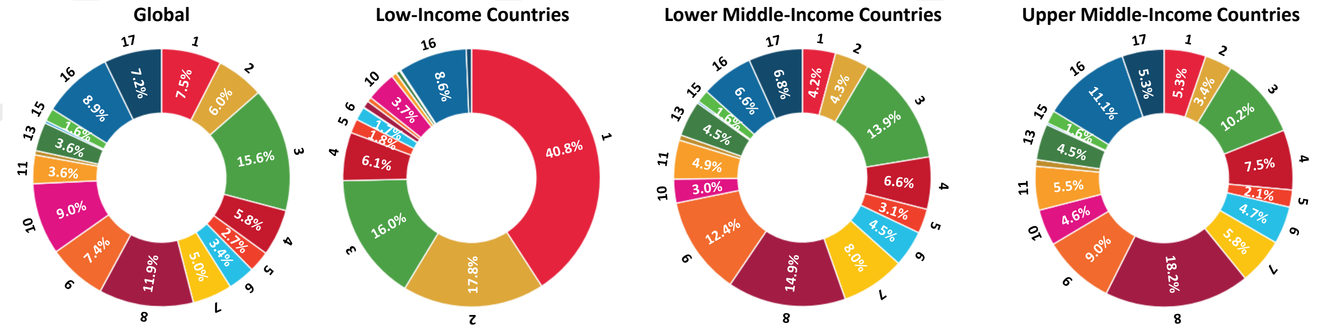

Figure 3: Distribution of Estimated SDG Budget Allocation by GNI per Capita. SDGs with a ratio of less than 1% have been omitted.

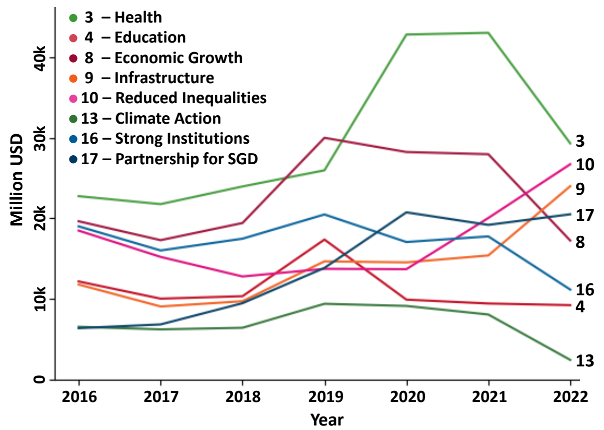

Figure 4: Estimated annual budget allocation. Over a 7-year period for a subset of SDGs with the most variability.

eight aid experts participated in the study, belonging to different government agencies. Further details about the experts are provided in the Appendix.

The interviews revealed three main challenges in the allocation of SDGs: categorization difficulties in multifaceted projects, variations in expertise and priorities between the personnel involved in the assignment, and the influence of policy priorities. Addressing these challenges requires nuanced project classification and standardized evaluation criteria.

Efficacy of the AI model could be summarized into three parts: efficiency, standardization, and potential. First, experts found that the use of AI model in assigning the aid goal to each project reduces time and increases the precision of their given task. For example, the multidimensional nature of development projects, especially those that target various objectives such as energy efficiency and climate change mitigation, has been adding complexities in assigning multiple SDGs. Experts identified that model recommendations can act as a guideline that reduces human labor and increases accuracy by giving more time to review the answers rather than solving them, indicated in the quote below:

> *"It could significantly reduce administrative workload, increasing efficiency, especially for those reviewing numerous projects annually. (OPC)"*

Second, experts found that our AI model can enhance credibility and coherence by minimizing human subjectivity and political influence, offering a standardized framework for aid classification and policy evaluation. This capability, in particular, addresses stakeholders' concerns about inconsistencies, arbitrariness in officials' interpretations, and the lack of clear guidelines, all of which threaten data quality, as quoted:

> *"Automatic SDG assignment provides a standardized guideline that reduces subjective variations and greatly helps establish uniformity. (KOICA)"*

Third, the model has the potential to support the efforts of the OECD and the UN to align aid statistics with the SDGs by encouraging broader participation in completing the currently optional *SDG focus index* field. By simplifying this process, it can promote wider participation among governments and strengthen their commitment to sustainable development.

5.3 Conclusion

We presented a model that uses LLM to classify international aid projects into SDG categories. This provides practical implications, assisting project managers in the time-consuming task of manually assigning multiple SDGs and supporting government experts in reviewing annual aid data. Our approach could potentially lead to a revision of the CRS guidelines, making SDG categorization mandatory and improving data quality and completeness, thus improving the reliability of development assessments.

We introduced three components to address the challenges in the classification of aid projects. Analytically, we compiled a new dataset that enriches existing aid information. The model shows potential for conducting detailed retroactive analyses of aid financing trends and donors' behavior, enabling a critical assessment of whether aid activities align with SDG commitments and meet recipients' demand.

The integration of AI in aligning aid projects with goal priorities offers significant theoretical implications. Our study serves as a salient example of how innovative technology can contribute to global governance for sustainable development, offering a foundation for further research. The expert interview also suggests that, by improving scalability, transparency, and aid effectiveness, our approach advances the SDG reporting system and informs future efforts to improve data-driven decision-making in international development.

Acknowledgments

This work was supported by the Center for Science Technology for Global Development (CSTG) at Korea Advanced Institute of Science and Technology (KAIST) Research Program (Grant Number: A0601006006) and the National Research Foundation of Korea (2022S1A5A8052217).

Contribution Statement

S. Park and D. Lee contributed equally as first authors and were primarily responsible for data analysis. All authors contributed to data interpretation and manuscript preparation. M. Cha and K.R. Park conceptualized the study and jointly serve as co-corresponding authors.

References

[Alsayyad, 2020] Amina Said Alsayyad. Linkages between official development assistance and the sustainable development goals: a scoping review. *International Policy Centre for Inclusive Growth*, page 446, 2020.

[Collier and Dollar, 2002] Paul Collier and David Dollar. Aid allocation and poverty reduction. *European Economic Review*, 46(8):1475–1500, 2002.

[Dalgaard *et al.*, 2004] Carl-Johan Dalgaard, Henrik Hansen, and Finn Tarp. On the empirics of foreign aid and growth. *Economic Journal*, 114:191–191, 02 2004.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Easterly and Pfutze, 2008] William Easterly and Tobias Pfutze. Where does the money go? best and worst practices in foreign aid. *Journal of Economic Perspectives*, 22(2):29–52, June 2008.

[Easterly, 2014] William Easterly. Are aid agencies improving? *Economic Policy*, 22(52):634–678, 08 2014.

[Ericsson and Mealy, 2019] Fredrik Ericsson and Sam Mealy. Connecting official development assistance and science technology and innovation for inclusive development: Measurement challenges from a Development Assistance Committee perspective. 2019.

[Gao *et al.*, 2021] T Gao, X Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proc. of the EMNLP*, 2021.

[Hynes and Scott, 2013] William Hynes and Simon Scott. The evolution of official development assistance: Achievements, criticisms and a way forward. 2013.

[Jang *et al.*, 2017] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *Proc. of ICLR*, 2017.

[Kasneci, 2018] Ola Kasneci. A results agenda for the 2030 agenda: New approaches for changing contexts. 2018.

[Knack *et al.*, 2011] Stephen Knack, F Halsey Rogers, and Nicholas Eubank. Aid quality and donor rankings. *World Development*, 39(11):1907–1917, 2011.

[Lee *et al.*, 2021] Eunji Lee, Sundong Kim, Sihyun Kim, Sungwon Park, Meeyoung Cha, Soyeon Jung, Suyoung Yang, Yeonsoo Choi, Sungdae Ji, Minsoo Song, et al. Classification of goods using text descriptions with sentences retrieval. *arXiv preprint arXiv:2111.01663*, 2021.

[Lee *et al.*, 2023] Junho Lee, Hyeonho Song, Dongjoon Lee, Sundong Kim, JiSoo Sim, Meeyoung Cha, and Kyung-Ryul Park. Machine learning driven aid classification for sustainable development. In *Proc. of the IJCAI*, pages 6040–6048, 2023.

[Ma *et al.*, 2021] Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. Label-specific dual graph neural network for multi-label text classification. In *Proc. of the ACL-IJCNLP*, pages 3855–3864, 2021.

[Morgenstern and Brown, 2022] Emily M Morgenstern and Nick M Brown. Foreign assistance: An introduction to US programs and policy. *Congressional Research Service*, 10, 2022.

[OECD, 2018] OECD. Proposal to include an SDG focus field in the CRS database. DCD/DAC/STAT(2018)1, 2018.

[Park and Park, 2024] Kyung Ryul Park and Young Shil Park. Addressing institutional challenges in sustainable development goals implementation: Lessons from the republic of korea. *Sustainable Development*, 32(1):1354–1369, 2024.

[Park *et al.*, 2022] Sungwon Park, Sundong Kim, and Meeyoung Cha. Knowledge sharing via domain adaptation in customs fraud detection. In *Proc. of the AAAI*, volume 36, pages 12062–12070, 2022.

[Pincet *et al.*, 2019] Arnaud Pincet, Shu Okabe, and Martin Pawelczyk. Linking aid to the sustainable development goals–A machine learning approach. 2019.

[Read *et al.*, 2009] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. In *In proc. of the ECML-PKDD*, pages 254–269. Springer, 2009.

[UNAIDS, 2004] UNAIDS. *Creditor Reporting System on Aid Activities Aid Activities in Support of HIV/AIDS Control Volume 2004 Issue 6: Activités d'aide pour la lutte contre le VIH/SIDA Volume 2004-6*. OECD publishing, 2004.

[Wang *et al.*, 2022] Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proc. of the ACL*, pages 7109–7119, 2022.

[Xiao *et al.*, 2019] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. Label-specific document representation for multi-label text classification. In *Proc. of the EMNLP-IJCNLP*, pages 466–475, 2019.

[Zhang and Chen, 2018] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in Neurips*, 31, 2018.