SDG Classification Using Instruction-Tuned LLMs

Tobias Fankhauser and **Simon Clematide** and **Martin Volk**
Department of Computational Linguistics
University of Zurich
tobias.fankhauser,simon.clematide,martin.volk@uzh.ch

Abstract

This paper investigates the potential of quantized, instruction-tuned Large Language Models (LLMs) for zero-shot classification of scientific abstracts according to the United Nations' Sustainable Development Goals (SDGs). We introduce the Decompose-Synthesize-Refine-Extract (DSRE) framework, leveraging advanced prompting techniques for both single-label and multi-label classification scenarios. DSRE is designed to enhance the zero-shot capabilities of LLMs for this domain-specific task. We explore the trade-offs between model performance and computational efficiency introduced by quantization. The performance of DSRE and quantized LLMs is benchmarked against fine-tuned LLM baselines and the Aurora system. Our findings demonstrate the potential of instruction-tuned LLMs for zero-shot SDG classification but emphasize the continued value of fine-tuning for optimal performance. Additionally, we consider dataset imbalance and the impact of augmenting datasets.

1 Introduction

In 2015, the United Nations outlined the 2030 Agenda for Sustainable Development, introducing 17 Sustainable Development Goals (SDGs) to address global challenges such as poverty, inequality, and climate change (United Nations: Department of Economic and Social Affairs). The academic community plays a vital role in advancing these goals through research that contributes to the targets defined within each SDG. However, effectively aligning scientific literature with specific SDGs remains a challenge due to the implicit nature of many contributions and the vast scope of subjects.

Recent advancements in Natural Language Processing (NLP), particularly in Large Language Models (LLMs) and generative AI, offer promising avenues for automating the classification of scientific abstracts into relevant SDGs.

This paper explores the application of quantized, instruction-tuned LLMs for zero-shot SDG classification of scientific abstracts. We propose the Decompose-Synthesize-Refine-Extract (DSRE) framework, an advanced prompt decomposition approach tailored for both single-label and multi-label classification tasks. DSRE aims to maximize the zero-shot capabilities of LLMs for domain-specific classification, addressing the need for data-efficient methods in aligning research outputs with SDGs.

We examine the balance between computational efficiency and model performance afforded by quantization and assess the effectiveness of DSRE compared to parameter-efficient fine-tuning approaches and the Aurora system (Vanderfeesten et al., 2022). Our investigation also covers the impact of dataset imbalance on classification accuracy and evaluates the potential of dataset augmentation strategies to mitigate these challenges.

2 Background and Related Work

SDG Classification and Aurora System: SDG classification addresses aligning scientific literature with the United Nations' Sustainable Development Goals (SDGs). Traditional methods have varied from rule-based systems, highlighting the need for manual keyword refinement (Rivest et al., 2021; Wang et al., 2023), to advanced transfer-based machine learning approaches. Among these, the Aurora system stands out for employing the multilingual BERT (mBERT) for SDG classification of scientific publications (Vanderfeesten et al., 2022). Aurora offers a binary classifier for each SDG, therefore supporting multilabel classification by setting a threshold.

LLMs and Instruction Tuning: Large Language Models (LLMs), such as GPT and BERT, have significantly advanced the field of NLP, offering deep contextual understanding and generation capabilities through extensive pretraining and fine-tuning processes (Radford et al., 2018; Devlin

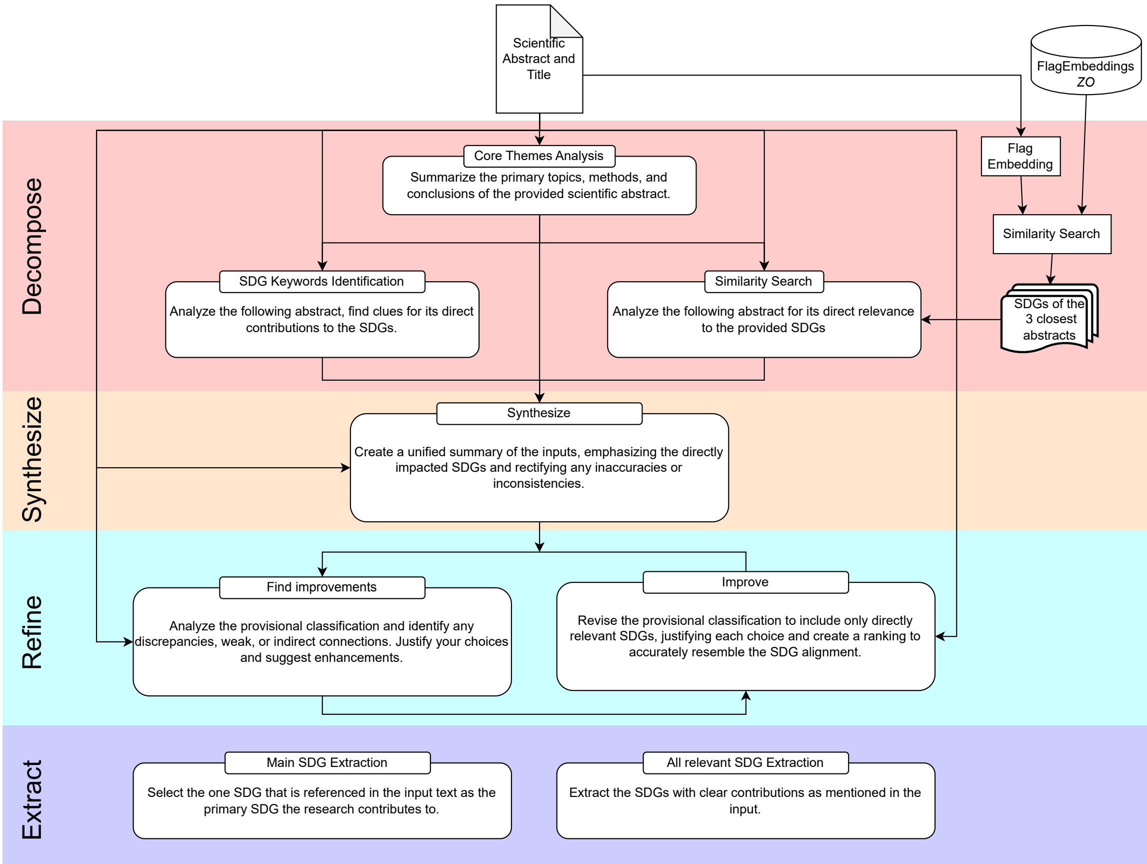

Figure 1: The DSRE prompt decomposition architecture using task decomposition (including semantic similarity), synthesis of results, self-refinement, and final SDG extraction.

et al., 2018). Instruction-tuned LLMs, leveraging datasets of instruction-output pairs, have shown versatility in adapting to specific tasks, without extensive task-specific training (Zhang et al., 2023).

Prompt Engineering techniques have emerged as critical for eliciting desired outputs from instruction-tuned models. These techniques range from manual and automated prompting to advanced methods like Chain-of-Thought (CoT) prompting, which guides models through intermediate reasoning steps (Wei et al., 2023), and In-Context-Learning (ICL) for task adaptation with minimal examples (Liu et al., 2023b).

Advanced Prompting Techniques: CARP, introduced by (Sun et al., 2023), directs the LLM to identify keywords and tones within the text. These are then assessed in a subsequent reasoning step, increasing text classification accuracy.

BSM, developed by (Saha et al., 2023), tackles complex problems by breaking them down into manageable subtasks. Each subtask is independently solved and the results are then merged to produce a coherent output. This method has shown to align LLM outputs with human-like reasoning.

Self-Refinement, proposed by (Madaan et al., 2023), enhances LLM outputs through an iterative process of feedback, and refinement. It continuously refines its responses based on self-generated feedback.

Parameter-Efficient Fine-tuning (PEFT): To address the computational demands of LLMs, techniques like Low Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) have been proposed. LoRA minimizes the number of parameters needed for fine-tuning by introducing low-rank matrices to the attention heads (Hu et al., 2021). QLoRA extends this by applying quantization, significantly reducing the computational and memory requirements, making fine-tuning feasible on consumer hardware (Dettmers et al., 2023).

Quantization methods enhance the efficiency of Large Language Models (LLMs) by reducing the precision of model weights from 32-bit floating points to lower precisions such as 4-bit, which decreases model size and computational demands while minimizing performance loss (Frantar et al., 2022). For the quantization mapping, *quantization constants* are derived from the values of each ten-

Model	Qu	A/R ↑	P	F1	RT	Mem
Zephyr 7B beta	-	**.608**	**.693**	**.615**	41.9	9.53
Zephyr 7B beta	Q	.570	.659	.570	**1.1**	8.73
Zephyr 7B beta	DQ	.551	.644	.547	1.2	**8.61**

Table 1: Performance comparison highlighting the effects of quantization (Qu) in zero-shot classification. No quantization (-), quantized (Q), and double quantized (DQ) models are compared. Accuracy (A), Recall (R), Precision (P),F1 Score (F1), Runtime in seconds (RT), and Memory usage in Gigabytes (Mem) are reported.

sor (Dettmers et al., 2023). These constants are used for restoring the weights to a higher precision, such as 16-bit, during inference and backpropagation in fine-tuning. Storing weights in 4-bit precision while performing computations in 16-bit during inference helps maintain numerical stability and reduces cumulative errors. This preserves the fidelity of the neural network operations even if the initial precision loss has already occurred. Despite the reduced precision, quantized models maintain essential functionalities such as contextual understanding and reasoning, as shown by Liu et al. (2023a).

3 Experiments

In our experiments, we test how the techniques described above influence SDG classification quality.

3.1 Data and Datasets

We use the imbalanced ZORA dataset *Z* (Meister, 2022), consisting of 384 scientific abstracts (270 train, 114 test items) from the Zurich Open Repository and Archive (ZORA). Each item belongs to a single SDG class (there is no class for not relevant to any SDG).

To compensate for our sparse in-domain data, we additionally exploit the crowdsourced OSDG Community Dataset (OSDG; UNDP IICPSD SDG AI Lab; PPMI), specifically selecting 26k texts (dataset label *O*) with good inter-annotator agreement (Roady, 2023). The dataset *ZO* simply concatenates *Z* and *O*, whereas *ZO Up* provides exactly 56 samples for each SDG (except for SDG 17, which is not present in *O*).

Aurora uses the dataset *A* stemming from Scopus abstracts retrieved by search queries for "Mapping Research Output to the Sustainable Development Goals (SDGs)" v5.0.2.[1]

[1]https://zenodo.org/records/4883250 (accessed on May 15, 2024)

Model	Qu	FT	A/R ↑	P	F1
Aurora	-	A	.500	.593	.500
LLAMA 2 7B	Q	-	.556	.626	.542
Zephyr 7B beta	Q	-	.570	.659	.570
DSRE	Q	-	.579	.687	.572
Zephyr 7B beta	-	-	.608	.693	.615
LLAMA 2 7B Chat	DQ	O	.653	.711	.647
LLAMA 2 7B Chat	DQ	Z	.657	.679	.653
LLAMA 2 7B Chat	DQ	ZO	.657	.709	.653
Zephyr 7B beta	DQ	Z	.685	.691	.679
LLAMA 2 7B Chat	DQ	ZO Up	.702	.710	.697
Zephyr 7B beta	DQ	ZO Up	**.711**	**.754**	**.717**

Table 2: Performance comparison on *Z* test set (N=114), highlighting the effects of quantization, prompt decomposition and fine-tuning. Column 'Qu' indicates whether no quantization (-), 4-bit quantization (Q) or double 4-bit quantization (DQ) was used. Column Fine-Tuning (FT) indicates which dataset was used to train the model, if any. The table is sorted by column Accuracy/Recall (A/R). We report mean values of weighted macro-averages per class from at least 2 runs.

3.2 Large Language Models

LLAMA 2 7B incorporates optimized multi-head attention mechanisms and employs strategies to enhance training efficiency (Touvron et al., 2023). It achieves new state-of-the-art results, with its 13B variant outperforming larger models like GPT-3 in Zero-Shot Common Sense Reasoning tasks. For our experiments, we use the smallest LLAMA 2 7B variant of this instruction-tuned LLM family.

LLAMA 2 7B Chat is an instruction-tuned chat variant trained using Reinforcement Learning with Human Feedback (RLHF) (Kaufmann et al., 2023). In our experiments, we always use the 4-bit double-quantized version for fine-tuning.[2]

Zephyr 7B beta is built by distillation techniques (Hinton et al., 2015) from the larger Mistral 7B teacher model (Tunstall et al., 2023). The so-called Distilled Direct Preference Optimization (dDPO) method adjusts the model's outputs to favor higher-quality responses in a supervised manner, thereby simplifying the alignment process and minimizing the reliance on reinforcement learning techniques or reward models.[3]

3.3 Quantization Experiments

The results in Table 1 show the expected loss in classification performance with each additional

[2]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf (accessed on May 15, 2024)

[3]https://huggingface.co/HuggingFaceH4/zephyr-7b-beta (accessed on May 15, 2024)

quantization level. However, if maximal performance is not key, runtime reduces by a factor of 40 with 4-bit quantization. Double quantization unfortunately reduces performance without tangible benefit in runtime or memory usage.

3.4 DSRE Experiments

The proposed Decompose-Synthesize-Refine-Extract (DSRE) approach aims to classify abstracts into SDGs by employing zero-shot learning via instruction-tuned LLMs (here always Zephyr 7B beta). This approach does not suffer from limited and imbalanced labeled data. Additionally, it can easily be formulated for multilabel predictions, which might be more adequate for scientific output that contributes to more than one goal.

DSRE is structured into four phases: Decomposition, Synthesis, Refinement, and Extraction. Figure 1 shows our architecture and the prompts.

The *Decomposition* breaks the classification task into smaller tasks and incorporates strategies akin to CARP, which enhance the model's ability to identify textual clues (Sun et al., 2023). Inspired by the similarity search applied in CARP, DSRE uses Flag Embeddings (BGE) within its Similarity Search component. Flag Embeddings leverage the last hidden state of the [CLS] token for sentence embeddings, providing dense vector representations of texts. Employing BGE allows DSRE to conduct semantic comparisons between the target abstract and a pre-classified corpus (Xiao et al., 2023).

The *Synthesis* merges these detailed analyses into a preliminary classification of the abstract's relevance to SDGs. This integration draws on methodologies related to BSM for synthesizing information from disparate sources (Saha et al., 2023).

Subsequently, the *Refinement* improves this preliminary classification by addressing inaccuracies and inconsistencies, drawing upon Self-Refinement techniques to improve the classification's accuracy and reliability (Madaan et al., 2023).

Lastly, the *Extraction* isolates the SDG classifications from the refined output. In our case, it requires fine-tuning with a specialized dataset to accurately extract SDG labels from DSRE outputs, indicating a specific limitation of output control from current LLMs.

3.5 Results

Table 2 compares 8 system configurations, providing further insights on the effects of quantization, advanced prompt engineering, and fine-tuning.

While our advanced DSRE prompting technique marginally improves classification accuracy within the group of quantized zero-shot approaches, it also introduces significant computational overhead (overall processing time per average abstract is around 46 seconds on an RTX 3080). Only exceptional classification performance would justify such an effort. However, the underlying Zephyr 7B beta is clearly not strong enough for zero-shot SDG classification reasoning. In the next section, we discuss the results of parameter-efficient fine-tuning experiments.

3.6 QLoRA Fine-Tuning Experiments

Both LLAMA 2 7B Chat and Zephyr 7B beta profit from fine-tuning. The lower part of Table 2 shows that a domain shift (and maybe also an annotation policy shift) from OSDG documents to scientific abstracts exists. For fine-tuning, carefully balancing classes by adding OSDG O material instead of concatenating all O data gives the best results. Parameter-efficient fine-tuning on in-domain training data overcomes the observed performance losses due to quantization by a large margin.

4 Conclusions

For the task of classifying short abstracts into Sustainable Development Goals (SDGs), like those typically encountered in ZORA, utilizing double-quantized Large Language Models (LLMs) via parameter-efficient fine-tuning methods proves to be a superior strategy. This approach notably excels in optimizing the use of computational resources and in reducing processing time. Furthermore, we expect our labelled data to grow within the ongoing project that this work is part of, and these models can immediately profit from this.

The DSRE method may offer an advantage when applied to longer full-text articles. The intermediary outputs generated by DSRE-like procedures could additionally serve to provide AI-generated explanations to end users, enhancing the interpretability of SDG classification decisions.

Future work will focus on deepening multilabel experiments that we already started on a small (N=51) dataset collected in a human annotation campaign using the output of the multilabel DSRE prompt as preannotations.

Acknowledgments This work was supported by the Digitalization Initiative of the Zurich Higher

Education Institutions (DIZH).

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

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

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2023. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2023a. Do emergent abilities exist in quantized large language models: An empirical study. *arXiv preprint arXiv:2307.08072*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Robin Meister. 2022. Identifying and prototyping a machine learning based SDG-classification for scientific papers. Bachelor's thesis, University of Zurich.

OSDG; UNDP IICPSD SDG AI Lab; PPMI. 2023. OSDG community dataset (OSDG-CD).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. electronic.

Maxime Rivest, Yury Kashnitsky, Alexandre Bédard-Vallée, David Campbell, Paul Khayat, Isabelle Labrosse, Henrique Pinheiro, Simon Provençal, Guillaume Roberge, and Chris James. 2021. Improving the Scopus and Aurora queries to identify research that supports the United Nations Sustainable Development Goals (SDGs) 2021. *Mendeley Data*, 2.

Jessica Saemi Roady. 2023. Automatic classification of academic papers according to the UN sustainable development goals – an interdisciplinary perspective. Master's thesis, University of Zurich.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

United Nations: Department of Economic and Social Affairs. 2021. The 17 goals: Sustainable development.

Maurice Vanderfeesten, Robert Jaworek, and Lennart Keßler. 2022. AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs). *Zenodo*.

Weiwei Wang, Weihao Kang, and Jingwen Mu. 2023. Mapping research to the Sustainable Development Goals (SDGs). *Research Square*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general Chinese embedding. *arXiv preprint arXiv:2309.07597*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

ZORA. 2024. Zurich open repository and archive.

A Appendix

A.1 QLoRA Configuration

```
BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=True,
    bnb_4bit_quant_type="nf4",
    bnb_4bit_compute_dtype=torch.bfloat16
)


LoraConfig(
    lora_alpha=16,
    lora_dropout=0.1,
    r=64,
    bias="none",
    task_type="CAUSAL_LM"
)


SFTConfig(
    per_device_train_batch_size=1,
    gradient_accumulation_steps=4,
    learning_rate=1e-4,
    logging_steps=10,
    num_training_epochs=1,
    bf16=True,
    optim="paged_adamw_8bit"
)
```

Listing 1: QLoRA configuration used for training

A.2 Fine-Tuning Prompts

```
The SDGs are:
SDG 1 – No Poverty: Aims to end poverty in all its forms
    everywhere.
SDG 2 – Zero Hunger: Aims to end hunger, achieve food
    security and improved nutrition, and promote
    sustainable agriculture.
SDG 3 – Good Health and Well–being: Aims to ensure
    healthy lives and promote well–being for all at all ages.

SDG 4 – Quality Education: Aims to ensure inclusive and
    equitable quality education and promote lifelong
    learning opportunities for all.
SDG 5 – Gender Equality: Aims to achieve gender equality
    and empower all women and girls.
SDG 6 – Clean Water and Sanitation: Aims to ensure
    availability and sustainable management of water and
    sanitation for all.
SDG 7 – Affordable and Clean Energy: Aims to ensure
    access to affordable, reliable, sustainable, and clean
    energy for all.
SDG 8 – Decent Work and Economic Growth: Aims to
    promote sustained, inclusive and sustainable economic
     growth, full and productive employment and decent
    work for all.
SDG 9 – Industry, Innovation and Infrastructure: Aims to
    build resilient infrastructure, promote inclusive and
    sustainable industrialization, and foster innovation.
SDG 10 – Reduced Inequalities: Aims to reduce inequality
    within and among countries.
SDG 11 – Sustainable Cities and Communities: Aims to
    make cities and human settlements inclusive, safe,
    resilient, and sustainable.
SDG 12 – Responsible Consumption and Production: Aims
    to ensure sustainable consumption and production
    patterns.
SDG 13 – Climate Action: Aims to take urgent action to
    combat climate change and its impacts.
SDG 14 – Life Below Water: Aims to conserve and
    sustainably use the oceans, seas, and marine resources
    for sustainable development.
SDG 15 – Life on Land: Aims to protect, restore and
    promote sustainable use of terrestrial ecosystems,
    sustainably manage forests, combat desertification, and
     halt and reverse land degradation and halt biodiversity
     loss.
SDG 16 – Peace and Justice Strong Institutions: Aims to
    promote peaceful and inclusive societies for
    sustainable development, provide access to justice for
    all and build effective, accountable and inclusive
    institutions at all levels.
SDG 17 – Partnerships for the Goals: Aims to strengthen the
     means of implementation and revitalize the global
    partnership for sustainable development.
SDG 0 – No Goal: Must be chosen if and only if the abstract
     does not contribute to any of the SDGs.
```

Listing 2: The list of SDG descriptions provided to the model

A.2.1 Prompt for Training and Inference

```
System Message:

You are an expert in scientific research, policy analysis and
    sustainable development. Determine to which
    Sustainable Development Goal (SDG) the paper
    described by the abstract contributes the most.\nThe
    SDGs are:\n{SDG_List}\n\nExpected Output Format:\
    nThe abstract primarily contributes to SDG [SDG
    Number] – [SDG Title].\n\nExample Output:\nThe
    abstract primarily contributes to SDG 1 – No Poverty


User Message:

Analyze the abstract and determine to which Sustainable
    Development Goal (SDG) the paper described by the
    abstract contributes the most.\nAbstract and Title:\n\"{
    Abstract_Text}\"
```

Listing 3: Prompt to train and infer the main SDG from an abstract

A.3 DSRE Prompts

```
General Constraints:

Always provide answers that are short, concise and to the
    point. Keep your answer within 300 words. Ensure
    clarity and consistency in your responses, avoiding
    contradictory statements or redundancy.
```

Listing 4: General Constraint provided to each Prompt, to keep the output short and concise.

A.3.1 Generation Prompt

```
Assess if the provided prompt meets the following criteria:

Clarity and Specificity: Is the prompt clear and specific
        enough to guide the model's response accurately?
Balancing Creativity and User Intent: Does the prompt find
        a balance between leveraging the model's creativity
        and achieving the user's intended outcome?
Domain Adaptation: Is the prompt tailored to a specific
        domain or context to enhance relevance and accuracy?
Handling Ambiguity: Does the prompt include strategies for
         effectively dealing with ambiguous inputs or
        situations?

Provided prompt:
"{System and user prompt}"
```

Listing 5: Prompt used to refine the prompts utilized in the DSRE using ChatGPT

A.3.2 Prompts for SDG Classification

```
System Message:

You are an AI expert in scientific abstract analysis. {
        general_constraint}


User Message:

Summarize the primary topics, methods, and conclusions of
        the provided scientific abstract. Abstract:\n\"{
        Abstract_Text}\"
```

Listing 6: Prompt utilized to extract core themes from the abstract

```
System Message:

You are an AI expert trained across multiple disciplines
        relevant to the Sustainable Development Goals (SDGs)
        , including environmental, social, economic, cultural,
        and technological fields. {general_constraint}
{sdg_description}

User Message:

Given the identified core themes:
"{Core_Themes}"

Analyze the following abstract for its direct relevance to the
        SDGs. Scientific Abstract and Title:
"{Abstract_Text}"
```

Listing 7: Prompt for analyzing abstracts for SDG relevance based on core themes

```
System Message:

You are an AI trained across multiple disciplines relevant to
        the Sustainable Development Goals (SDGs), including
        environmental, social, economic, cultural,
        technological fields, and scientific analysis. Critically
        and objectively analyze texts for their contribution to
        Sustainable Development Goals (SDGs). Summarize
        the main goal of the provided SDG and assess whether
         the abstract directly contributes towards achieving this
         SDG. Justify instances where the abstract does not
        directly contribute to the SDG. {general_constraint}

User Message:

Given the identified core themes:
"{Core_Themes}"

Analyze the following scientific abstract for its direct
        relevance to the SDGs "{closest_abstract_sdgs}, {
        closest_sdg_centroid}". Scientific Abstract and Title:
"{Abstract_Text}"
```

Listing 8: Prompt for detailed SDG relevance analysis of scientific abstracts

```
System Message:

You are an AI specialized in synthesizing and summarizing
        complex texts related to the same scientific abstract.
        Your task is to merge various inputs into a single,
        coherent summary. Focus specifically on the directly
        impacted Sustainable Development Goals (SDGs).
        Exclude irrelevant or indirectly affected SDGs, and
        correct any inconsistencies or inaccuracies in the
        information. All inputs provided are aspects of the
        same scientific abstract. {general_constraint}

User Message:

Here are the texts to be summarized, all pertaining to the
        same scientific abstract:

1. Scientific Abstract and Title:
"{Abstract_Text}"

2. Core Themes of the Scientific Abstract:
"{Core_Themes}"

3. SDG Concepts of the Scientific Abstract:
"{SDG_Concepts}"

4. Additional Information about the Scientific Abstract:
"{Retrieval}"

Create a unified summary of these inputs, emphasizing the
        directly impacted SDGs and rectifying any
        inaccuracies or inconsistencies.
```

Listing 9: Prompt for merging inputs into a coherent summary with a focus on SDGs

```
System Message:

You are an AI trained in critical analysis and sustainable
        development goals (SDGs). Your task is to critically
        analyze the provided abstract against the provisional
        classification for alignment with relevant SDGs.
        Highlight any discrepancies, weak, or indirect
        connections. {general_constraint} Refer to the
        provided SDG descriptions for accurate comparisons.

{sdg_description}

{general_constraint}

User Message:

Scientific Abstract and Title:
"{Abstract_Text}"
```

```
Provisional Classification:
"{Response_Text}"

Analyze the provisional classification and identify any
        discrepancies, weak, or indirect connections. Justify
        your choices and provide enhancements.
```

Listing 10: Prompt for finding improvement in the provisional classification

```
System Message:

You are an AI tasked with optimizing classifications related
        to Sustainable Development Goals (SDGs). Revise the
        provided provisional classification by addressing the
        identified enhancements. Ensure that your revised
        classification only includes directly relevant SDGs and
        clearly justifies the direct relevance of each selected
        SDG. Adjust the rankings and explanations to reflect a
        more accurate alignment with the SDGs.

{general_constraint}

User Message:

Scientific Abstract and Title:
"{Abstract_Text}"

Provisional Classification:
"{Defective_Response_Text}"

Identified Enhancements:
"{Identified_Issues_Text}"

Revise the provisional classification to include only directly
        relevant SDGs, justifying each choice and create a
        ranking to accurately resemble the SDG alignment.
```

Listing 11: Prompt for optimizing classifications related to SDGs

```
System Message:

As a precise AI, your specific function is to identify the
        single, most pertinent Sustainable Development Goal (
        SDG) from those mentioned in the input in relation to
        the abstract. Select the one SDG that is referenced in
        the input text as the primary SDG the research
        contributes to based on the content of the abstract.
        Your response must be formatted as 'SDG X', where 'X'
        is the number of the most relevant SDG from those
        mentioned in the input. Ensure your response strictly
        adheres to this format and excludes any justification or
        additional information.

User Message:

Input:
"{BSM_Response}"
```

Listing 12: Prompt for identifying the primary SDG from input

```
System Message:

Analyze the provided input to identify which Sustainable
        Development Goals (SDGs) are directly impacted by
        the scientific abstract. Extract only the SDGs with
        clear contributions as mentioned in the input. Present
        your findings as a list of SDG numbers, separated by
        commas. Your response should be only this list,
        without additional justification or explanation.

User Message:

Input:
"{BSM_Response}"
```

Listing 13: Prompt for identifying all directly impacted SDGs from input

A.4 Extraction Module Fine-Tuning

A.4.1 Example Training Input-Output Pair

```
Input Text:

Ranked List of SDG Contributions:

1. SDG 15 – Life on Land: This study's main contribution is
        toward SDG 15, which aims to protect terrestrial
        ecosystems and promote sustainable use of resources.
        Accurately estimating population trends for amphibian
        species is crucial for understanding which species are
        declining and which require conservation action. By
        accounting for imperfect detection, this study helps
        identify species at risk of decline and promotes
        conservation efforts to protect their habitats.

2. SDG 17 – Partnerships for the Goals: The study
        emphasizes the importance of collaboration and data
        sharing, highlighting the need for partnerships to
        inform conservation policy and manage species
        declines. By advocating for the routine recording of
        nonobservation and absence data, the study
        underscores the significance of partnerships and
        collaborative efforts for conservation purposes.

3. SDG 13 – Climate Action: While not directly related to
        climate action, this study's findings regarding species
        declines could have implications for understanding the
        impacts of climate change on amphibian populations.
        Accurate population trend estimates are necessary for
        understanding which species are declining and which
        are most in need of conservation action, and this study'
        s insights could inform climate adaptation and
        mitigation strategies for amphibian species.

The study does not contribute significantly to the other
        SDGs and should not be ranked for them. If the
        abstract did not contribute to any of the SDGs, SDG 0
        – No Goal – would be assigned.


In conclusion, this study's contributions are primarily
        aligned with SDG 15 and SDG 17, emphasizing the
        significance of protecting terrestrial ecosystems and
        promoting partnerships and collaboration for
        conservation purposes. Its insights also have indirect
        implications for SDG 13, highlighting the need for
        accurate population trend estimates to understand the
        impacts of climate change on amphibian populations.
```

Primary SDG Extraction Output:

SDG 15

Secondary SDGs Extraction Output:

15, 17, 13

Listing 14: Input-output pairs for fine-tuning the DSRE extraction Module

A.5 Artifacts Availability

The artifacts associated with this research can be found in the repository on
https://github.com/TobiFank/SDG-Classification-Using-Instruction-Tuned-LLMs/blob/master/README.md