

# Integrating Expert Labels into LLM-based Emission Goal Detection: Example Selection vs Automatic Prompt Design

Marco Wrzalik<sup>1</sup>, Adrian Ulges<sup>1</sup>, Anne Uersfeld<sup>1</sup>, Florian Faust<sup>2</sup>, Viola Campos<sup>1</sup>

<sup>1</sup> RheinMain University of Applied Sciences (Wiesbaden, Germany)

<sup>2</sup> Sociovestix Labs Ltd. (Kaiserslautern, Germany)

<sup>1</sup>firstname.lastname@hs-rm.de, <sup>2</sup>firstname.lastname@sociovestix.com

## Abstract

We address the detection of emission reduction goals in corporate reports, an important task for monitoring companies' progress in addressing climate change. Specifically, we focus on the issue of integrating expert feedback in the form of labeled example passages into LLM-based pipelines, and compare the two strategies of (1) a dynamic selection of few-shot examples and (2) the automatic optimization of the prompt by the LLM itself.

Our findings on a public dataset of 769 climate-related passages from real-world business reports indicate that automatic prompt optimization is the superior approach, while combining both methods provides only limited benefit. Qualitative results indicate that optimized prompts do indeed capture many intricacies of the targeted emission goal extraction task.

## 1 Introduction

The urgency of the climate crisis necessitates immediate action across various sectors of the global economy. These efforts are targeted at *Net Zero*, i.e. achieving a balance between emitted and removed greenhouse gases, particularly CO<sub>2</sub>. Policies such as the European Union's *Green Taxonomy* aim to redirect financial investment flows toward sustainable businesses, setting incentives for companies committing to concrete emission reduction goals.

To evaluate and monitor these commitments, analysts must review extensive corporate documentation, including annual reports, sustainability reports, and stewardship disclosures. This manual process of locating and extracting relevant data, often referred to as *carbon accounting*, is labor-intensive and time-consuming. Identifying such climate goals is a surprisingly intricate task, as analysts have to distinguish concrete, binding, self-imposed and quantitative emission goals from vague statements, greenwashing, externally mandated requirements, goals *related to* climate change

(such as "moving out of coal"), etc. Overall, the task not only requires high precision but also poses unique challenges due to the nuanced language and diverse reporting styles used by companies.

To increase the efficiency of carbon accounting, Large Language Models (LLMs) have emerged as powerful tools. Given a *prompt* consisting of task-specific instructions and a text passage to analyze, the LLM outputs whether the passage contains an emission goal. When integrating such LLMs into practical workflows, their outputs are inspected and – in case of errors – corrected by analysts, resulting in a set of labeled "challenge samples" that grows over time. These examples offer an interesting option for *in-context learning* (ICL), i.e. to improve the LLM's accuracy by utilizing the examples to improve the prompt, without applying fine-tuning to the model.

In this paper, we investigate ICL strategies to improve LLMs with expert knowledge in form of labeled examples, focusing on the task of emission goal detection. Particularly, we compare two approaches: (1) *Example selection*, which incorporates a limited number of few-shot examples into the prompt. These exemplify the desired behavior, and are selected dynamically to resemble the input passage. (2) *Automatic Prompt Design*, in which the LLM adjusts its own instructions. This approach applies an iterative optimization process in which erroneous challenge cases are inspected, reflected, and new, refined instructions are generated and evaluated.

While both techniques hold promise, their effectiveness in real-world applications has not been compared extensively. We hope to fill this gap with the following contributions:

1. We conduct a comprehensive comparison of example selection and automatic prompt design on the task of emission goal classification, using 769 passages from the public *NetZero-Facts* dataset (Wrzalik et al., 2024).

2. We highlight the strengths and limitations of each prompting strategy, guiding practitioners applying LLMs to sustainability classification tasks. Specifically: (1) example selection improves results, especially with weak prompts; (2) auto-prompting yields greater overall gains; and (3) in automatic prompt design, few-shot examples provide limited additional benefits.

## 2 Related Work

This section reviews existing literature on prompting strategies that make use of labeled data, focusing on example selection for few-shot learning and automatic prompt design.

### 2.1 Example Selection for Few-Shot Learning

Liu et al. (2022) investigated the selection of good in-context examples for GPT-3 and found that semantically similar examples, chosen based on proximity in embedding space, significantly improve model performance. Rubin et al. (2022) proposed a contrastive learning-based method that learns to retrieve task-specific examples, showing significant performance improvements by optimizing the input-output pairs for correct predictions. Su et al. (2023) introduced a selective annotation framework that enhances few-shot learning by strategically selecting diverse and representative examples from a small annotated pool. Their graph-based approach demonstrates that such careful example selection can lead to significant performance improvements across various natural language processing tasks. Zhang et al. (2022) framed example selection as a reinforcement learning problem, proposing a Q-learning-based approach to actively select examples. This method shifts from similarity-based retrieval to learning a policy that optimizes example selection for few-shot learning, demonstrating moderate improvements on downstream tasks. In this work, we focus on the selection of semantically similar examples as motivated by Liu et al. (2022).

### 2.2 Automatic Prompt Design

The task of optimizing prompts for LLMs has received increasing attention, with several approaches leveraging the model’s ability to self-improve: Shin et al. (2020) introduced *AutoPrompt* uses a gradient-guided search to iteratively refine trigger tokens, resulting in competitive performance compared to manually designed prompts

for tasks like natural language inference. Zhou et al. (2023) proposed *Automatic Prompt Engineer* (APE), which iterates over LLM-generated candidate prompts, selecting and refining those that lead to the best performance based on task-specific score functions, often outperforming human-generated prompts. Similarly, Pryzant et al. (2023) introduced *ProTeGi*, a method that optimizes prompts based on LLM-generated reasonings over erroneous examples. The performance of these reasonings is estimated and steers a beam search that explores multiple candidate options. Yang et al. (2024) utilize the LLM itself as a general-purpose optimizer capable of refining prompts and solving various tasks. Their method, OPRO, iteratively generates and evaluates prompts based on the history of previous attempts, framing prompt optimization as a general meta-optimization task. Finally, *Intent-based Prompt Calibration* (IPC) by Levi et al. (2024) introduces the auto-generation of synthetic challenge cases to calibrate prompts, which are then labeled by the expert. Our work explores automatic prompt design – as outlined above – from a practitioners perspective. Specifically, we follow a simplified variant of Pryzant et al. (2023)’s *ProTeGi*, which (instead of beam search) performs a greedy search.

### 2.3 Emission Information Extraction

The detection and extraction of SDG (Sustainable Development Goals) related information has recently gained traction as a research field: Spokoyny et al. (2023) bridge the gap between structured reporting (in form of questionnaires) and unstructured reporting in form of text: Their ClimaBench benchmark challenges NLP models to extract climate-related information from company reports and thus auto-fill questionnaires. Schimanski et al. (2024) pretrain specific BERT models for environmental, social and governance aspects on 13.8 million corporate disclosures and curate three balanced 2k-document test collections for the E, S and G dimensions. Their textual ESG scores explain a substantial share of the variance in leading commercial ratings, demonstrating that domain-tuned NLP can markedly narrow the long-standing "rating gap" in ESG measurement. These two works do not address recent large-scale LLMs, and thus focus on task-specific fine-tuning instead of in-context-learning (as investigated in this work).

Other, more recent approaches tackle SDG information extraction with LLMs: Garigliotti

Table 1: Results of Few-shot Prompting (%), including example selection in Row 4 ("Similar").

EXAMPLES	simple instructions				expert instructions			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Zero-shot	56.2	52.2	100.0	68.6	81.2	77.1	86.1	81.4
Static	72.3	63.8	97.0	77.0	82.0	90.8	69.4	78.7
Random	66.2	58.9	98.6	73.7	80.5	75.7	89.5	81.5
Similar	69.8	62.3	92.9	74.6	87.0	87.2	85.4	86.3

(2024) investigate an LLM-based RAG pipeline, which – similar to our approach – addresses the task of detecting textual passages containing Sustainable Developmental Goal (SDG) targets. Thulke et al. (2024) introduce *ClimateGPT*, a 7B–70B-parameter, climate-specialised LLM family pretrained on 4.2B domain tokens and instruction-tuned with expert tasks, matching Llama-2-70B on bespoke climate benchmarks while reducing hallucinations through retrieval augmentation. While both approaches are based on LLMs, they use manual prompt tuning and do not address in-context learning from user feedback, which is the focus of our work.

### 3 Approach

Given a text passage from a sustainability report, we classify whether this passage contains a relevant emission goal. To achieve this, we prompt an instruction fine-tuned LLM, specifically OpenAI’s GPT-4o mini. The prompt includes both a set of task-specific instructions and – optionally – a small number of few-shot examples, i.e. input/output pairs that demonstrate the desired behavior. For a passage containing a relevant emission goal, the LLM’s answer is expected to be "True," while irrelevant passages should be classified "False." Within this framework, we investigate two methods to include expert knowledge in form of labeled examples: *Few-Shot Example Selection* and *Automatic Prompt Design*.

#### 3.1 Few-shot Example Selection

Here, we add labeled few-shot examples to the prompt, which are selected to be semantically similar to the target passage. For example, the sentence “*Our goal is to reduce paper waste in our administrative departments by 35% by 2027*” might be misclassified as a relevant goal. However, injecting a similar example with the correct label such as “*By 2028, we aim to reduce paper usage in our offices by 50% through digitalization*” into the prompt arguably improves the chance of a correct result.

We embed passages using *Sentence Transformer* (Reimers and Gurevych, 2019), specifically *all-MiniLM-L6-v2*, which balances efficiency and performance in MTEB (Muennighoff et al., 2023). This model is fine-tuned on 1.17 billion sentence pairs (Reimers, 2024) using contrastive learning. We retrieve similar passages via cosine similarity in the embedding space.

Since class distribution affects the LLM’s decision, we select up to three examples per class, prioritizing the most similar ones to form five few-shot demonstrations.

#### 3.2 Automatic Prompt Design

Given the training set of labeled passages, we apply an approach similar to *ProTeGi* (Pryzant et al. (2023)): Starting from an initial prompt, the LLM iterates through the training set of labeled text passages in random order. Given a text passage  $p$ , the current prompt – consisting of instructions  $\mathcal{I}$  and optionally some static few-shot demonstrations – is used to predict whether  $p$  contains a relevant emission goal. If the model’s prediction contradicts the ground truth, we prompt the LLM to analyze possible root causes of the error. This step is inspired by the *Chain-of-Thought* method proposed by Wei et al. (2022). We then feed the LLM’s rationale back to the LLM, prompting it to modify  $\mathcal{I}$  to correct the error so that the desired label is generated in future predictions. We obtain a new prompt candidate, comprising of new instructions  $\mathcal{I}'$  (and optionally the same few-shot examples).

The accuracy of this new prompt with instructions  $\mathcal{I}'$  is assessed by computing its F1 score on the training set. Should this exceed the score of the previous instructions  $\mathcal{I}$  by at least a small margin  $\epsilon$ , the new prompt is accepted and the iteration is continued with  $\mathcal{I}'$ . The margin  $\epsilon$  ensures an observed improvement to be statistically significant, and also limits the instruction complexity (we found most modifications made by the LLM to add new clauses and/or sentences to the instruction). In contrast to the work by Pryzant et al. (2023), we do not em-

ploy beam search but a simple greedy search. This is to limit the computational cost associated with high numbers of evaluations.

## 4 Experiments

We compare the effectiveness of the above approaches for example selection and automatic instruction design in improving emission goal detection with LLM prompting.

**Dataset** We use the NetZeroFact-BIG Dataset (Wrzalik et al., 2024), which contains passages from 16 business reports labeled by analysts. A passage is relevant (true) if it yields a correct fact; otherwise, it is false.

The data has been split into a test set (on which we estimate performance metrics) and a training set (from which we draw few-shot examples in example selection, and which the automatic prompt design iterates over). To prevent data leakage between the splits due to duplicate statements within the same report, we split along the reports: Four reports have been chosen whose 207 passages form the test set, while 562 passages from the remaining 12 reports form the training set. 36%/48% of labels in the training/test split are positive.

**Setup and Technical Details** We have run all experiments for two versions of the starting instructions  $\mathcal{I}^1$ : (1) A *simple* ad-hoc version, and (2) an *expert* version, which resulted from a manual process of iterative optimization and result inspection prior to the experiments presented in this paper. As few-shot examples, we either use none ("Zero-shot"), 5 random ones from the training set ("Random"), a fixed set of 5 examples, which were expert-selected to be particularly informative prior to our experiments ("Static"), or example selection as described above ("Similar").

We report well-known quality metrics, namely classification accuracy, precision, recall and the F1 score. Our experiments were run with *GPT-4o mini*<sup>2</sup> through the *OpenAI API* (OpenAI, 2024) with Python’s *LangChain* framework.

We set the margin to  $\epsilon=0.01$ , which corresponds approximately to two times the standard deviation observed in many of our experiments (despite greedy generation, we found OpenAI’s output to be non-deterministic, which is why we repeated

evaluation 7 times and report the average). With this margin, we found the prompt tuning process to converge in less than one epoch.

**Few-Shot Example Selection** Table 1 displays test results for simple (left) and expert (right) prompts with the different few-shot example selection strategies. Note that only few-shot examples (and not the prompt) are varied in this experiment. We see that adding few-shot examples consistently improves results for the simple prompt, with expert-selected examples ("Static") complementing this prompt best. For the expert prompt, however, only the similarity-based example selection yields an improvement by  $\approx 5\%$ . We hypothesize that this complements the – already quite elaborate – prompt best. Also, note that few-shot prompting improves precision rather than recall (likely because examples emphasize intricacies of the extraction task).

**Automatic Prompt Design** Table 2 illustrates results for automatic prompt design. First, and most importantly, we observe strong improvements over the manual prompts in the zero-shot case (Rows 1+2), from 68.6% to 88.2% for the simple prompt and from 81.4% to 86.9% for the expert prompt (note that – surprisingly – starting the optimization process from the simple prompt works even better). In both cases, automatic prompt design outperforms example selection. The Appendix gives a qualitative impression of the evolution of the prompt, outlining which aspects were added in the optimization process, and that – indeed – the resulting prompt reflects on some key intricacies of the extraction task.

Few-shot examples in this setting deteriorates results consistently when no examples were used in prompt tuning (Row 2 vs. Rows 3-5). It seems that these examples confuse the model with its highly specific instructions. When using static examples in training<sup>3</sup>, we observe improvements in some cases but not with example selection (last row).

## 5 Conclusion

Overall, our results suggest that – particularly with ad-hoc prompts, in which little explicit knowledge of the extraction task is encoded – automatic prompt design is more effective compared to example selection. A surprising finding is that, when

<sup>1</sup>All instructions and examples can be found in the appendix.

<sup>2</sup>The version used is gpt-4o-mini-2024-07-18

<sup>3</sup>For efficiency reasons, we stick with static examples, since fixed prompts during prompt tuning allow for batching.



Table 2: Results of Automatic Prompt Design (%).

EXAMPLES DURING...		simple start instruction				expert start instruction			
Tuning	Testing	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
<i>(no tuning, zero-shot)</i>		56.2	52.2	100.0	68.6	81.2	77.1	86.1	81.4
Zero-shot	Zero-shot	88.1	84.2	92.5	88.2	86.9	83.3	90.8	86.9
Zero-shot	Static	83.2	80.4	85.9	83.0	87.0	92.2	79.7	85.4
Zero-shot	Random	82.3	75.7	93.9	83.6	84.9	83.1	85.9	84.3
Zero-shot	Similar	83.1	80.3	85.7	82.9	84.9	87.0	80.4	83.6
Static	Zero-shot	87.4	90.1	82.7	86.2	89.9	88.2	90.9	89.5
Static	Static	89.9	88.7	90.5	89.6	89.1	93.5	83.0	87.9
Static	Random	81.7	76.0	91.3	82.8	89.0	88.0	89.9	88.6
Static	Similar	84.1	84.2	82.3	83.2	87.4	90.5	82.4	86.3

applying automatic prompt design, we found few-shot samples, particularly when drawn with different strategies compared to prompt tuning, to be harmful in some cases.

Future research on the issue may include experiments with other LLMs (particularly open-source or open-weight ones), the extension to other sustainability-oriented information extraction tasks (such as reported de-facto emissions, which often come in tables), and methods for interactive prompt-codesign by expert and LLM.

## Limitations

One key limitation of our study is that we only focus on OpenAI’s o4-mini model, such that – also due to the rapid advancement of large language models (LLMs) – our findings merely represent a snapshot at the time of our experiments. While we conducted initial evaluations with more recent LLaMA-3 models (Grattafiori et al., 2024) and observed similar performance trends as for the OpenAI-based experiments in this paper, findings may differ for newer models with enhanced reasoning capabilities such as DeepSeek-R1 (DeepSeek-AI et al., 2025). Specifically, these latest models have been claimed to offer improved performance on tasks requiring long reasoning chains. It should be, however, that our specific task (emission goal detection) relies more on the precise assessment of edge cases rather than extended reasoning, suggesting that our core findings remain relevant despite these developments. Future work should systematically evaluate newer models to assess their potential impact on this task.

Another limitation of our study is the focus on company reports as the only – and inherently non-objective – source of information. While our extraction methods focus on identifying verifiable facts,

the lack of external validation poses a risk of bias in the results. Here, a valuable direction for future research would be to cross-match extracted facts with independent sources, such as social media discussions, reports from non-governmental organizations (NGOs), or investigative journalism. This could provide a more comprehensive and balanced assessment of corporate emissions goals.

## 6 Acknowledgements

This work was supported by the German Federal Ministry of Education and Research (BMBF), Project "NetZeroNet" (01IS22050).

## References

- DeepSeek-AI, Daya Guo, Dejian Yang, et al. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). arXiv:2501.12948.
- Dario Garigliotti. 2024. [SDG target detection in environmental reports using retrieval-augmented generation with LLMs](#). In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 241–250, Bangkok, Thailand. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. [The Llama 3 Herd of Models](#). arXiv:2407.21783.
- Elad Levi, Eli Brosh, and Matan Friedmann. 2024. [Intent-based prompt calibration: Enhancing prompt optimization with synthetic boundary cases](#). *CoRR*, abs/2402.03099.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What Makes Good In-Context Examples for GPT-3?](#) In *Proc. DeeLIO@ACL 2022, May 27, 2022*, pages 100–114.

- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. **MTEB: Massive Text Embedding Benchmark**. In *Proc. EACL 2023, May 2-6, 2023*, pages 2006–2029.
- OpenAI. 2024. The most powerful platform for building AI products. <https://openai.com/api/>. Accessed: 2024-10-21.
- Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. **Automatic Prompt Optimization with "Gradient Descent" and Beam Search**. In *Proc. EMNLP 2023*, pages 7957–7968.
- Nils Reimers. 2024. Huggingface models: sentence-transformers/all-minilm-l6-v2. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>. Accessed: 2024-10-08.
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. In *Proc. EMNLP 2019*.
- Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. **Learning To Retrieve Prompts for In-Context Learning**. In *Proc. NAACL 2022, July 10-15, 2022*, pages 2655–2671.
- Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in esg measurement: Using nlp to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.
- Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. **AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts**. In *Proc. EMNLP 2020, November 16-20, 2020*, pages 4222–4235.
- Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. **Towards answering climate questionnaires from unstructured climate reports**.
- Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. **Selective Annotation Makes Language Models Better Few-Shot Learners**. In *Proc. ICLR 2023, May 1-5, 2023*.
- David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. **Climategpt: Towards AI synthesizing interdisciplinary research on climate change**. *CoRR*, abs/2401.09646.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. In *Proc. NeurIPS 2022*.
- Marco Wrzalik, Florian Faust, Simon Sieber, and Adrian Ulges. 2024. **NetZeroFacts: Two-Stage Emission Information Extraction from Company Reports**. In *Proc. FinNLP Workshop @ LREC-COLING 2024*, pages 70–84.
- Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024. **Large Language Models as Optimizers**. In *Proc. ICLR 2024, May 7-11, 2024*. OpenReview.net.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. **Active Example Selection for In-Context Learning**. In *Proc. EMNLP 2022, December 7-11, 2022*, pages 9134–9148.
- Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. **Large Language Models are Human-Level Prompt Engineers**. In *Proc. ICLR 2023, May 1-5, 2023*. OpenReview.net.

## Appendix

### A Prompts

#### A.1 Simple Instruction

System: Determine if the text describes a commitment to reducing carbon emissions or achieving net zero; return "True" if it does, otherwise return "False".  
 Human: <example input>  
 AI: <example output>  
 ...  
 Human: <input passage>  
 AI: <output prediction>

#### A.2 Expert Instruction

System: You are an information extraction tool for climate goals that classifies whether a given text contains a statement about the commitment to a goal regarding carbon emissions. I will present you with passages from asset managers' reports. You will determine whether the given text contains a commitment to either a specific relative reduction in carbon emissions or to achieving net zero or carbon neutrality. Ignore any vague statements; a target is only a target if it states by when the target is to be achieved. For relative emission reductions, a specific percentage reduction must be stated. Ignore goals of third parties. Your answer is 'True' if the statement contains such a climate target of the asset manager and 'False' if it does not.  
 Human: <example input>  
 AI: <example output>  
 ...  
 Human: <input passage>  
 AI: <output prediction>

### A.3 Static Few-Shot Examples

System: <Instruction>  
Human: A standout feature of the GreenTech Solutions Factory is its pledge to function entirely on renewable energy sources, aiming for net-zero emissions across its operations.  
AI: False  
Human: We are participating in the UN-backed Net-Zero Asset Owner Alliance (AOA) where a large number of the worlds biggest investors commit themselves to being net carbon neutral in their investments by no later than 2050 and to continually make five-year sub-targets for CO2 footprints.  
AI: True  
Human: Net zero emissions means achieving a balance between greenhouse gas (GHG) emissions produced and the amount removed from the atmosphere, consistent with limiting global warming to 1.5C and neutralising the impact of any residual emissions by permanently removing an equivalent amount of carbon dioxide (CO2). For BTPS this will mean reducing the portfolio's emissions through changing investments and investing in technologies which reduce emissions.  
AI: False  
Human: We have also emphasised our green ambitions by announcing that, from 2019-2025, we will reduce the carbon footprints of our investments by 29 per cent.  
AI: True  
Human: In our climate targets published in 2019, we are committed to exiting from investments in thermal coal by 2025. We are also committed to excluding oil exploration from our investments by 2030.  
AI: False  
Human: <input passage>  
AI: <output prediction>

### A.4 Automatic Prompt Tuning: Prompt for Instruction Modification

System: <instruction>  
Human: <example input>  
AI: <example output>  
...  
Human: <input passage>  
AI: <output prediction>  
Human: Your prediction is wrong, we expect the answer to be "<target label>". Analyze the error in your previous prediction and identify which part of the instruction may have led to the incorrect result.  
AI: <output reasoning>  
Human: Modify the instruction to improve understanding of the expected behavior for the next prediction. Ensure the modification corrects the mistake without reversing the labels. Provide only the modified instruction in your response.  
AI: <output instruction>

### A.5 Interaction Example from Automatic Prompt Engineering

The following is a dialogue from the Automatic Prompt Engineering process, starting with the simple instruction and not including the few-shot examples during optimization.

System: Determine if the text describes a commitment to reducing carbon emissions, achieving net zero, or setting specific emission reduction targets; return "True" if it does, otherwise return "False".  
Human: 2 Guide for Identifying Sustainable Financing. 3 Identified Staff is made up of directors, senior managers or employees whose professional activities have a significant impact on the risk profile of an entity. An environmental and climate strategy that aims to contribute to the sustainable transition, addressing the challenge of accelerating the transition to a carbon neutral economy, taking into account the natural capital.  
AI: True  
Human: Your prediction is wrong, we expect the answer to be "False". Analyze the error in your previous prediction and identify which part of the instruction may have led to the incorrect result.  
AI: Upon reevaluating the text, it does not explicitly mention a commitment to reducing carbon emissions, achieving net zero, or setting specific emission reduction targets. Instead, it discusses an environmental and climate strategy aimed at contributing to a sustainable transition towards a carbon-neutral economy, which is more about the broader goal rather than specific commitments or targets.  
Human: The error in my previous prediction stemmed from interpreting the phrase "contribute to the sustainable transition" as a commitment, when it does not specify concrete actions or targets related to emission reductions. The instruction to identify specific commitments or targets was not adequately applied in this case.  
Human: Therefore, the correct answer should indeed be "False." Thank you for your feedback.  
Human: Modify the instruction to improve understanding of the expected behavior for the next prediction. Ensure the modification corrects the mistake without reversing the labels. Provide only the modified instruction in your response.  
AI: Determine if the text explicitly describes a commitment to reducing carbon emissions, achieving net zero, or setting specific, measurable emission reduction targets. Return "True" if it does, otherwise return "False." Focus on clear statements of intent or quantifiable goals rather than general strategies or aspirations.

## B Automatic Prompt Tuning: Instruction Evolution

### B.1 Initial Prompt: Simple Instruction - Zero-Shot

**Initial Instruction:** *Determine if the text describes a commitment to reducing carbon emissions or achieving net zero; return "True" if it does, otherwise return "False".*

#### Summarized Instruction Evolution:

1. Emphasizes *specific* emission reduction targets.
2. Focus on explicit statements of intent or measurable targets.
3. Excludes vague or general strategies.
4. Adds requirement for numerical values or timelines in commitments.
5. Stresses actionable targets with clear timelines.
6. Focuses on specific commitments, not just data or ambitions.
7. Clarifies that numerical values alone are insufficient.
8. Emphasizes intent to achieve targets within a timeframe.
9. Notes need for long-term commitment or endpoint.
10. Recognizes specific phrasing like "reduce to zero by 2050."

**Tuned Instruction:** *Determine if the text explicitly describes a commitment to reducing carbon emissions, achieving net zero, or setting specific, measurable emission reduction targets with clear numerical values or timelines. Return "True" if it does, otherwise return "False." Focus on clear, quantifiable statements of intent or goals, ensuring that any mentioned targets are specific, actionable, and include a defined timeline or endpoint. Emphasize that a commitment must include a direct statement of intent to achieve a specific target within a specified timeframe, rather than general goals or investments in technology without clear emission reduction outcomes. Note that numerical data or reduction ambitions alone, without a clear long-term commitment or endpoint, do not qualify as a commitment. Additionally, ensure to recognize that phrases indicating intent to achieve specific targets by a certain year (e.g., "reduce to zero by 2050") are valid commitments.*

### B.2 Initial Prompt: Manually Tuned Instruction - Zero-Shot

**Initial Instruction:** *You are an information extraction tool for climate goals that classifies whether a given text contains a statement about the commitment to a goal regarding carbon emissions. I will present you with passages from asset managers' reports. You will determine whether the given text contains a commitment to either a specific relative reduction in carbon emissions or to achieving net zero or carbon neutrality. Ignore any vague statements; a target is only a target if it states by when the target is to be achieved. For relative emission reductions, a specific percentage reduction must be stated. Ignore goals of third parties. Your answer is 'True' if the statement contains such a climate target of the asset manager and 'False' if it does not.*

#### Summarized Instruction Evolution:

1. Adds emphasis on the need for both a clearly stated percentage and a target year for relative reductions, and a target year for net zero/carbon neutrality.
2. Stresses the validity of the target only if it explicitly states the target year for both relative reductions and net zero/carbon neutrality. Includes checks to ensure the presence of a target year for all goals.
3. Refines further by requiring that net zero or carbon neutrality commitments specify a particular year (e.g., 2030), not just a vague timeframe. Reinforced the need for specific target years and percentages for all commitments.

**Tuned Instruction:** *You are an information extraction tool for climate goals that classifies whether a given text contains a statement about the commitment to a goal regarding carbon emissions. I will present you with passages from asset managers' reports. You will determine whether the given text contains a commitment to either a specific relative reduction in carbon emissions (with a clearly stated percentage and a target year) or to achieving net zero or carbon neutrality (with a specified target year). A target is only valid if it explicitly states by when the target is to be achieved, including a specific year (e.g., 2030) for net zero or carbon neutrality commitments. For relative emission reductions, a specific percentage reduction must be stated along with a target year. Ignore goals of third parties and any vague statements. Your answer is 'True' if the statement contains such a climate target of the asset manager and 'False' if it does not. Ensure to check for the presence of a clearly specified target year for both relative reductions and net zero or carbon neutrality commitments.*