Towards Just Policy:

Identifying Distributive Justice Principles in a Global Climate Policy Context

S.J. van Santen

TU Delft

Towards Just Policy:

Identifying Distributive Justice Principles in a Global Climate Policy Context

By

S.J. van Santen

To obtain the degree in
Master of Science
in Engineering & Policy Analysis

at the Delft University of Technology,

Student number:	4660102	
Chair:	Prof. dr. mr. ir. N. Doorn	TU Delft
First Supervisor:	Dr. J. Zatarain Salazar	TU Delft
Second Supervisor:	Dr. N. Goyal	TU Delft
Daily Supervisor:	P. Biswas	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.
Models and data files are available at https://github.com/svansanten/TextToDistributiveJusticeMT

TU Delft Delft University of Technology

Preface

I would like to express my gratitude towards all the people who played a crucial role in the completion of this thesis. First, I want to thank all my supervisors for their time and enthusiasm. Our discussions have helped me greatly to shape this research and have motivated me by highlighting the importance of its findings. I explicitly want to thank Palok, with whom I could share my frustrations and whose scientific curiosity was contagious.

To all my friends, family, and especially Sam, thank you for your support and words of wisdom.

With this thesis, my time as a student comes to an end. I am the most grateful for all memories and lifelong friendships that I will cherish. I hope my work can serve as a foundation for future research and I am curious to see what the future holds.

Suze
Amsterdam, June 2024

Executive Summary

Global climate policymaking is complex and influenced by injustices and uncertainties. Designing policies for global climate mitigation have different consequences for all countries, both economical and societal. This leads to normative uncertainties, situations where different morally defensible courses of action can create divergent views on how to distribute resources, responsibilities, and risks. However, policies need to be perceived just by all countries in order to be truly effective. These diverging views are dependent on context and are ultimately based on moral rules and principles that prescribe when a distribution is morally just; distributive justice principles.

These principles are not only important during climate negotiations, as they can cause policy deadlocks, but also in development of policy evaluation models like Integrated Assessment Models. These models are subject to developers' choices as to what types of policies and whose interests to include, often assuming a single arbitrary distributive justice principle. Consequently, they can present compromised outcomes. A deep understanding of these principles and moral justifications is necessary to be able to account for them in policy and modelling research. This understanding can be developed with a bottom-up approach, evaluating stakeholders' views, or a top-down approach, focusing on moral theories. Both approaches have their limitations. The top-down approach is subject to discrepancy between theory and practice. With the bottom-up approach, it is often difficult to determine stakeholders' normative stance by simply asking them. Evaluation of texts presented by stakeholders for the use of distributive justice principles is a version of a bottom-up approach with theoretical foundations.

This research, in collaboration with TU Delft's HIPPO lab, uses Natural Language Processing (NLP), by means of OpenAI's GPT-4o model, to perform qualitative coding on High-Level Segment (HLS) speeches from the UNFCCC Conferences of the Parties. Model selection is based on the proven performance of OpenAI's GPT's and its increasingly widespread use in qualitative research. The unprompted nature of the High-Level Segment speeches makes them a compelling case for identifying moral justifications for distributions and policy preferences. Each speech is delivered in a consistent context, with all parties given equal time to speak during the same conference.

The study contributes to climate justice and integrated assessment modelling by offering theoretical insights and practical implications for both policymakers and model developers. By examining distributive justice perspectives in climate negotiations, it highlights the moral foundations of climate policy decisions and provides a valuable dataset for future justice-focused text analysis models. This research aims to answer the following research question:

To what extent do Large Language Models accurately identify distributive justice preferences in climate negotiation texts?

The research follows a workflow designed for LLM augmented text annotation. The process includes two parts:

1. Codebook and dataset creation: Distributive justice principles are defined in the context of climate policymaking. These definitions are used to deductively create a codebook, which is inductively updated during a sentences-based manual annotation of a subset of HLS speeches.
2. Model Evaluation: The codebooks are used instruct GPT-4o to classify sentences. Model performance is compared to the manual annotations by means of accuracy, precision, recall, and F1 scores.

Five key principles have been identified and defined in the context of climate policy. **Egalitarianism** advocates for reducing inequalities and ensuring fairness in the distribution of resources, opportunities, and responsibilities. **Utilitarianism** prioritizes maximizing overall welfare or benefits, disregarding individual, or subgroup differences. **Prioritarianism** emphasizes increasing the welfare

of the worst off, aiming to uplift those in the most disadvantaged positions**. Sufficientarianism** focuses on ensuring that everyone reaches a minimum threshold of well-being, while **libertarianism** emphasizes individual freedom and opposes forced redistribution.

A total of 51 HLS speeches are manually annotated, a subset of the full HLS speech corpus created. The first annotation step of annotation is determining relevance. Within this pre-processing task, a sentence is deemed relevant when it presents a normative statement. Next the relevant sentences are annotated for the motivational distributive justice principle and three additional categories that can be used to evaluate the context in which principles are identified. The first category is Topic, evaluating the main topic of the sentence. The other two categories refer to the practical implementation of the principle: Unit and Shape. Ultimately, five codebooks are created, one for each category and its corresponding labels.

The manual annotation revealed that labelling for distributive justice principle is a complex process where edge cases and annotator interpretations are highly influential. Large class imbalances are found, especially for the libertarian and sufficientarian principle. The ground truth dataset includes 267 relevant sentences, accounting for 17% of the total 1543 sentences. The prioritarian principle is most dominant and found in 72 sentences. Aggregating per country, the utilitarian is dominant in 15 of 51 speeches.

GPT-4o performance is evaluated in four annotation tasks; (1) pre-processing for relevant sentences, identifying distributive justice principles in both the full dataset (2) and pre-processed dataset (3), and identifying not only the principles, but also the additional categories Topic, Unit, and Shape (4). Instructions for the execution of these tasks are provided in prompts which are based on the created codebooks. For each task the benefit of adding an example is evaluated, comparing zero-shot and one-shot prompts. Overall, annotation of relevant sentences reached 60% accuracy, exceeding 70% for certain principles with one-shot prompts. Without pre-processing, the model showed high precision for not-relevant sentences but less accuracy for principles, often mislabelling not-relevant sentences. Despite good precision in some areas, the model struggled with overall accuracy and consistency compared to human annotations. However, misclassifications are found in more ambiguous labels which are also complex for human annotators to evaluate.

Evaluation of 41 speeches with GPT-4o took less than 20 minutes or about 30 seconds per speech. This is a fraction of manual annotation time of 20 minutes per speech. The model's performance underlines its potential, both for pre-processing and identification purposes. Additionally, costs are limited as all experiments were performed for a total of less than 75 dollars. However, the non-deterministic nature of the model combined with the large number of edge cases lead to inconsistencies in appointed labels, which must be re-evaluated by a human. Additionally, the black box nature of the model potentially leads to biases currently unknown. Using these models for classification tasks that are fundamentally normative can give a false sense of objectivity.

The generalizability of the results generated with GPT-4o, trained on an unspecified dataset, is limited to this model and may not directly apply to other generative LLMs. However, the theoretical foundations and provided definitions in the research can facilitate additional training. The ground truth dataset can be used to train other LLMs for supervised sentence-based classification tasks that can identify principles in climate related texts.

The study's findings emphasize the importance of a nuanced, bottom-up understanding of distributive justice principles in climate negotiations, revealing moral plurality and the limitations of focusing on a single principle. The identified practical implementation characteristics, such as Shape and Unit, provide deeper insights into policy preferences. This research contributes to climate justice and integrated assessment modelling, offering a foundation for future justice-focused text analysis models. The ground truth dataset and complete HLS speeches dataset are publicly available on GitHub, facilitating further research and training of other LLMs for supervised classification tasks.

Contents

Acronyms

COP – Conference of the Parties
UNFCCC – United Nations Framework Convention on Climate Change
HLS – High Level Segment
CBDR – Common but Differentiated Responsibilities and Respective Capabilities
NC – National Communications
NDC – Nationally Determined Contributions
LDC – Least Developed Country
IAM – Integrated Assessment Model
LLM – Large Language Model
API – Application Programming Interface
GPT – Generative Pre-trained Transformer
GNS – General normative statement
P – Precision
R – Recall
S - Support

1 | Introduction

In every decision involving some type of distribution - whether it is a distribution of food, water or other more complex elements like risk or responsibility – there are diverging views on how this distribution should look. These diverse perspectives can lead to normative uncertainties: situations in which there are different options or courses of action which can all be morally defensible (Taebi et al., 2021, Van Uffelen et al., 2024).

Normative uncertainties are evident in several contexts, but they are particularly influential in global climate policymaking. This type of policymaking is complex and characterized by policy deadlocks (Bagozzi, 2015; Lamb et al., 2020). One cause of complexity is found in the injustices of climate change: the people least responsible for carbon emissions, mainly living in developing countries, are the most vulnerable to the impacts of climate change (Chancel and Piketty, 2015; Füssel, 2010). Simultaneously, the measures taken to limit climate change can further disadvantage these people, potentially increasing existing social and economic inequalities (Islam and Winkel, 2017; Newell et al., 2021). This means that the costs and benefits of climate change are inherently distributed unequally.

Global climate policies need to be deemed just by all countries for them to be truly effective (Lamb et al., 2020). However, this notion of justice is very broad, referring to distributions, procedures, or recognition (Van Uffelen et al., 2024). It relates not only to the policymaking process but also the global distribution of the impacts that these policies might have on both society and the economy (Newell et al., 2021). Some deem a policy to be fair if every country needs to reduce their emissions an equal amount, others might deem this unfair and point to the fact that consequences of this policy would limit economic development. These diverging views are dependent on what is distributed and are ultimately based on moral rules and principles that prescribe when a distribution is morally just; distributive justice principles (Davidson, 2021, Van Uffelen et al., 2024)).

With the development of policy modelling, researchers have become more and more able to anticipate the effects of policies before they are implemented. In a climate policymaking context, policymakers and researchers use Integrated Assessment Models (IAMs) to anticipate the effects of policies under different climate scenarios (Weyant, 2017). These models include socioeconomic, technological, and biochemical variables, and aim to provide a quantitative description of the interactions between the earth and human system (Sahoo and Murari, 2023). However, these models are too influenced, either consciously or unconsciously, by normative uncertainties as they are subject to developers' choices as to what types of policies and whose interests to include (Jafino et al. 2021).

A famous example of a IAM is the DICE model, developed by Nordhaus (2018). This model provides an aggregated representation of climate change, mostly on a global level, and is developed with the use of simplified economic and climate dynamics. It can be used to optimize policies by balancing the marginal costs of emission reductions against the marginal costs of damages. Aggregation functions called social welfare functions (SWFs) are used to capture multiple performance metrics into a single overarching metric representing global welfare (Sikdar, 2009). This function can represent different types of distributive justice principles, with a utilitarian SWF disregarding distributions and maximizing overall welfare, prioritarian SWF favouring an increase in the benefit of the worse-off, and others attaching equity weights to individuals proportional to, for example, their consumption (Jafino et al., 2021).

The principles of distributive justice can be seen as moral guidelines, evaluate the dimensions of the distribution and determine if it is just. The adaptation of a different principle can lead to varying policy preferences. This is evident in IAMs that are used for mitigation planning, where the choice of principles in allocating mitigation budgets and aggregate welfare affect the distributional outcomes across nations (Adler et al., 2017). The utilitarian principle of distributive justice is most used in IAMs, mostly inexplicitly and its use is not often questioned (Kolstad et al. 2014; Rubiano Rivadeneira and

Carton, 2022; Zimm et al., 2024). This arbitrary use of distributional principles has been criticized and reduces the transparency of models (Jafino, 2021, Kartha et al., 2018). Where multiple principles of distributive justice can be found in real-life policy- making, IAMs generally aggregate values and base their evaluations on the same principle for all actors involved. They include complex variables, but currently have limited application for the evaluation of distributive justice (Jafino et al., 2021).

Dealing with normative uncertainties and diverging ideas of distributive justice in climate policymaking requires a deep understanding of the various perspectives and moral justifications used in different decisions and proposals (Taebi et al., 2020). This understanding can be generated from the bottom-up, evaluating the diverse views and experiences of stakeholders involved in a decision-making process or top-down, focussing on moral theories as guiding frameworks. Where the top-down approach leads to foundational understanding with limited practical application, the bottom-up approach presents more real-life insights that are less generalizable to decision-making contexts, but more practical. It is not always possible to comprehensively ask stakeholders for their moral stances, leading to the need to derive these perspectives elsewhere.

Van Uffelen et al. (2024) and Okekere (2010) highlight that principles of distributive justice can be articulated in policy documents, policy proposals and policy debates. There are vast amounts of documentations available that can be used to study the real-life use of distributive justice principles. The field of Natural Language Processing enables the processing of human language and presents a solution to this volume challenge. It provides tools and techniques that can be used to assess large text-based datasets (Sietsma et al., 2024; Stede and Patz, 2021). These techniques are increasingly used to evaluate climate-related text and United Nations Framework Convention on Climate (UNFCCC) documentation (Bagozzi, 2015; Webersinke et al. 2022; Wright et al., 2023). This research presents a theory driven application of a large language model to identify distributive justice principles evoked by stakeholders in policymaking discourse. Specifically, this study analyses the high-level segment speeches delivered at the UNFCCC Conference of the Parties (COP) with OpenAI's GPT-4o.

Conducted in collaboration with the TU Delft HIPPO lab, this research contributes the fields of climate justice and modelling. It offers both theoretical insights and practical implications for model developers and policymakers by deepening the understanding of distributive justice. By examining distributive justice perspectives in climate negotiations, it sheds light on the moral underpinnings of climate policy decisions. Additionally, this research will present a labelled dataset of distributive justice examples from UNFCCC discourse, relevant for future development of more justice-focused text analysis models.

Climate change is considered a grand challenge considering its global scale, inescapable impacts, and complex nature. Addressing the causes and consequences of climate change requires coordinated efforts between governments, industries, and individuals as well as interdisciplinary solutions from scientific fields like economics, biochemistry, and social science. Coordinated efforts call for consensus on climate policies, something that is difficult to achieve when stakeholders have diverging interests. This study's objective contributes to the understanding of these diverging interests and can be linked to policy recommendations on what approaches can be seen as promising for achieving consensus. Ultimately, the research touches on the fundamental elements of the EPA program; system understanding, modelling, policy and politics.

1.1 | Case introduction: UNFCCC COP High Level Segment Speeches

The UNFCCC produces various and numerous types of documentation, ranging from press releases and meeting reports to statements and final agreements. Other documents are provided by the parties themselves as the UNFCCC requires them to provide documentation on their efforts to reduce emissions, highlighting their national priorities and policy plans (UNFCCC, 1992). UNFCCC documents have been studied for various reasons, most of them focusing on national policies and specific issues found in NDCs and NCs (Biesbroek et al. 2022; Wright et al. 2023). Bagozzi (2015)

aims to determine national governments' foreign policy intentions regarding global climate issues with the use of topic modelling. He highlights the negotiation complexity of the UNFCCC COP, evaluating the High-level Segment (HLS) speeches at the 16th to 19th UNFCCC COP arguing that these speeches or statements correspond to countries' initial climate change positions in the debate.

All COP sessions to date have included a High-Level segment attended by high level members of the UN, governments, and heads of state. Among other things, it aims to facilitate agreement between parties on the major political issues and provides policy guidance by raising issues and signalling priorities (UNFCCC, 2020). During this segment, all 198 parties[1] have the opportunity to present their national statements. Within these speeches, parties outline their contributions to solving the climate change problem, as well as highlighting topics of concern and preferences for future policies. Parties can touch upon global distribution of resources, for example advocating for increased funding of developing countries, presenting their motivations for a global public. Bagozzi (2015) argues that these speeches present a unique and relatively unprompted window into the positions of parties towards climate change, prior to any concessions made during the negotiations.

Existing studies on the moral justifications in climate policies have largely concentrated on a limited number of countries or country-specific policies (e.g. Lykkeskov and Gjerris, 2017; Pottier et al., 2017). In contrast, analysing these high-level speeches from the COP offers a broader and more comprehensive understanding of global climate policy discourse, as they capture a wide array of perspectives and positions from different countries before any negotiations lead to concessions.

In conclusion, the unprompted nature of the High-Level Segment speeches makes them a compelling case for identifying moral justifications for distributions and policy preferences. Each speech is delivered in a consistent context, with all parties given equal time to speak during the same conference. Although not mandatory, most parties publicly publish their speeches in the UNFCCC database. By focusing on the last ten COPs (COP19 (2013) - COP28 (2023)) and evaluating only English-written texts, this study aims to provide a comprehensive and nuanced understanding of global climate policy discourse, capturing a wide array of perspectives and positions from different countries before negotiations lead to concessions.

1.2 | Research questions

This research aims to explore the potential use of LLMs for studying the use of distributive justice principles in texts, specifically focusing on HLS speeches from UNFCCC COPs. It has two main purposes: descriptive knowledge generation and prescriptive insights for both modelers and policymakers.

The main research question is:

To what extent do Large Language Models accurately identify distributive justice preferences in climate negotiation texts?

This research question is supported by four sub-questions. Each sub-question is evaluated separately and ultimately combined in the conclusion of this research.

1. What are the key theories and concepts of distributive justice relevant in the context of climate negotiations?

The concept of distributive justice is broad and can be defined in various ways. For the aim of this research, it is crucial to reach an understanding of distributive justice in relation to climate policymaking, as well as its main principles. A literature review is performed aiming to combine

[1] As of 2022, the UNFCCC has 198 parties. This includes all members of the UN and non-member observers (the state of Palestine and the Holy Sea). The European Union is seen as an individual party (UNFCCC, 2022). From this point on, the terms "party" and "country" are used interchangeably.

insights from social sciences, political philosophy, and climate studies. The conceptualization is used as the foundation for the identification stage of the research.

2. Which distributive justice principles and preferences are expressed by parties in HLS speeches?

The analysis of HLS speeches serves as a case study for assessing the capability of LLMs in annotating texts for distributive justice preferences. In assessing capabilities, it is crucial to compare labels presented by an LLM with a human annotated dataset (Pangakis et al., 2023). To create this human annotated dataset, a subset of speeches will be evaluated based on the theoretical framework derived in SQ1. This analysis will consist of combined inductive and deductive coding, meaning that codes and themes will emerge directly from the text, but are also informed by the definitions of distributive justice. The goal is to create a comprehensive codebook, which will detail the specific labels and themes relevant to distributive justice as expressed in the COP HLS speeches. This approach ensures that the coding framework remains flexible and accurately reflects the content of the speeches. The subset of speeches will be evaluated using this codebook, resulting in a manually annotated dataset.

3. How does LLM identification of distributive justice preference compare to human annotated data?

This research will use and evaluate the annotating capabilities of Large Language Models. LLMs are specific models in the NLP domain that use a deep learning approach and are trained on vast amounts of unlabelled textual data (Webersinke et al. 2022). These models have extensive capabilities and offer capabilities like creative writing, programming, and can be used as chatbots (Xiao et al., 2023). This research employs the most recent version of OpenAI's GPT-4; GPT-4o (OpenAI, 2024, Achiam et al., 2023). It will be applied for a quantitative observational analysis of distributive justice principles for the case-specific documents.

As noted by Pangakis et. al (2023), conceptual difficulties and text data idiosyncrasies can influence the performance of LLMs in annotation tasks. This leads to the need to validate the labels presented by the LLM against labels generated by humans. The labels created in sub-question 2 is used as ground truth. The performance of the LLM will be evaluated for the metrics accuracy, precision, recall and F1 score over repeated annotation tasks.

4. What are the strengths and limitations of LLM application for the distributive justice identification task?

The final sub-question used the insights gained from the LLM annotation to determine the main strengths and weaknesses found. Performance per annotation task is compared to identify strengths, and both technical and fundamental limitations. These can be related to general accuracy, but also generalizability and benchmarking against manual annotation. These insights can not only be used in relation to this specific case but also carries implications for broader applications across different types of texts.

1.3 | Report outline

The report's outline follows the sub-questions sequentially. Chapter 2 presents the conceptualization of distributive justice and its principles. In addition, a connection is made between the theory and the role of distributive justice in the UNFCCC. Chapter 3 elaborates on the workflow used in this research, consisting of two main parts. Chapter 4 describes the manual annotation process and the formulation of the codebooks. In Chapter 5, GPT-4o is used to carry out the same annotation tasks as performed during the manual annotation process. Chapter 6 highlights strengths and limitations of the research. Chapter 7 answers the research question and discusses this research's societal and scientific implications.

2 | Conceptualization

This chapter aims to answer the first sub-question, conceptualizing distributive justice in the context of climate policymaking. First a definition is formulated, followed by an explanation of the main principles found in literature. Next, an overview is presented of the role of Distributive Justice in the UNFCCC. Finally, a summary presents the main takeaways.

2.1 | Definition of Distributive Justice

The measures required to mitigate climate change come with many challenges and justice implications (Kolstad et al., 2014). In dealing with climate change, numerous justice related issues arise. These issues are fundamental, comprising of questions of how to deal with climate change, focussing on risks, rights, responsibilities, and distributions. This has led to debates on climate justice in both society and policymaking (Newell et al., 2021). The multifaced nature of the debate itself gives room to varying interpretations of climate justice in both policymaking and research (Schlosberg and Collins, 2014).

Various studies describe forms or dimensions that are a part of climate justice, most commonly naming procedural and distributive justice (Newell et al., 2021; Pottier et al., 2017; Zimm et al., 2024). Procedural justice relates to the decision-making process surrounding the impacts and responses to climate change. It accounts for the need for these procedures to be fair, accountable, and transparent (Newell et al., 2021). Distributive justice relates to the distribution of goods, benefits, and burdens among members of society that should be done in a just manner (Newell et al., 2021; Pottier et al., 2017; Zimm et al., 2024). It determines who receives what and provides reasoning why. Both procedural and distributive justice are important in climate change decision-making and can be seen as interrelated with procedural inequalities leading to unfair allocation of resources (Pottier et al., 2017).

Distributive justice itself can be divided into three dimensions: (1) Unit, or what is to be distributed, (2) Shape, the preferred shape of the distribution, and (3) Scope, determining between whom the distribution is made (Jafino et al., 2021; Newell et al., 2021). The scope implies the recipients of the distribution. They can be individuals or groups or include non-human actors. It considers a temporal aspect, determining the inclusion of future generations or solely focusing on the group currently living. In this regard, the scale of justice is also considered, with distributions applied to local, national, regional, or global scale (Van Uffelen et al., 2024). An example would be the distribution of income. This can be done on national scale or locally, for example in a city. It can account for all individuals at that scale, or a select group, such as low-income households. The temporal aspect is found in the duration of the redistribution, if it is only focused on current low-income households, or also taking long-term measures into account.

The unit can imply, for example, a distribution of money, welfare, jobs, opportunities, or risks (Lamont and Favor, 2017). In climate policymaking, distributions are mainly discussed in relation to mitigation[2] and adaptation[3] and its related costs. In this context, a different type of distribution can be highlighted: the distribution of climate responsibilities. This is related to the responsibility to meet global climate goals, often referring to taking implementing policies or bearing the costs of mitigation and adaptation measures (Caney, 2021). Another type of distribution found in the climate debate is the transfer of technology (Okekere, 2010). Technological solutions can help with both mitigation and adaptation and one can be of view that these should be accessible for everyone.

[2] Mitigation involves reducing the flow of greenhouse gases to the atmosphere. It involves reducing and avoiding emissions (EEA, 2024).

[3] Adaptation involves making adjustments to ecological, social or economic systems in response to the effects of climate change. An example of adaptation is the construction of buildings that can cope with extreme heat. (Caney, 2021)

The shape determines the pattern of the distribution and thus the preferred outcome of a policy. Referring to the redistribution of income; the shape could prescribe an equal amount of money to all individuals (equality-based shape) or differentiate based on circumstances like age or type of job (equity-based shape). Other examples are shapes that add more weight to individuals that have the least of a resource or based on individual needs, for example an investment needed for the implementation of a specific policy.

The dimensions provide information on how distributive justice is implemented in practice, covering who is affected, what is being distributed, and how the distribution occurs. They provide contextual information on what is seen as important and can be used to identify moral foundations on which opinions on distributive justice are used as well as indicating how distributive justice should be implemented in new policy. An example is the shape, where a distribution that maximizes utility for all is motivated by a utilitarian principle (Konow, 2003).

2.2 | Principles of Distributive Justice

Distributive justice principles present motivations for preferring specific distributions and are closely related to philosophical and ethical concepts. Their main difference is found in the prescription on what basis a just distribution should be made (e.g. based on maximization or equality). This research focusses on five different distributive justice principles. They have been selected because of their overlapping use in both climate modelling and climate policymaking (e.g. in: Zimm et al., 2024, Davidson, 2021; Jafino et al., 2021,, Kolstad et al. 2014; Meyer & Roser, 2006;).

2.2.1 | Egalitarian

The principle of egalitarianism favours distributions in which a greater degree of equality is reached than currently exists. Egalitarian justice rest on the fundamental premise that all persons have the same fundamental worth and dignity which commands respect. At its core, it advocates for a more equitable distribution of resources, opportunities, and burdens, with the overarching goal of reaching a fairer and more just society (Arneson, 2023).

Strict egalitarianism is one of the simplest principles of distributive justice. It is justified on the grounds that all people are morally equal and that allocating every person exactly the same amount is the fairest distribution (Lamont and Favor, 2017). Strict egalitarianism values the existence of all inequalities negatively and only refers to the outcome of the distribution. In climate policy, strict egalitarianism can be related to the idea that every country must take on action to contribute to solving the climate crisis. However, in practical terms, achieving strict equality is often challenging and may not be desirable. Setting an equal income level for all would mean that some are benefitted, and others are worse off. All would be equal in monetary terms, this does not mean that overall inequality, referring to welfare or women rights, is not accounted for. The equal distribution of climate action does not specify the level of action needed to be taken, where it can be deemed unjust to let a country with minimal emissions take the same actions as the countries with the highest emissions.

Broader interpretations of egalitarianism acknowledge the complexities of human society and do not completely disregard inequalities. When considering equality of opportunity, inequalities in outcomes can be justified, if “we start from a level playing field and everyone has equal opportunities, regardless of their background or circumstances” (Robeyns, 2019).

Applying equality of opportunity to climate policy means recognizing differences but ensuring all have an equal chance to achieve the same goals. The notion of fair climate policy mostly refers to not further increasing the existing inequalities and acknowledging the fact that there are inequalities (Arneson, 2023). This implies the need to take the circumstances of individuals or countries into account when making distributions, as it would not be fair for the richest and poorest to invest the same amount of money. This would decrease the chances of the poorest and potentially having only little effect on the richest. The same can be said for an egalitarian distribution of responsibilities.

Within this research, the broad definition of egalitarianism will be used. This indicates distributions referring to both strictly equal distributions as well as distributions aiming to reduce inequalities are seen as egalitarian.

2.2.2 | Utilitarian

The utilitarian principle prescribes an approach that determines moral rightness entirely on the consequences of an act. It does not take any notion of inequalities and the aim of utilitarianism is to maximize the benefits of all, indicating that the action that produces the most 'good' is the morally right one (Driver, 2022). These benefits or goods or welfare are aggregated, only evaluating the final sum. Every individual counts the same in this distribution, and differences between individuals or subgroups are disregarded. Following the Utilitarian principle, distributions are preferred that maximize the total sum of welfare.

In relation to climate policymaking, a preference for the utilitarian principle would be evoked in distributions that increase the global welfare, for example referring to country level investments that disregard any local preferences. In the distribution of responsibility or actions, the utilitarian principle can also be prescribing a reason as to why a distribution is preferred; for example, referencing the need to act for 'the global good'. This research also focuses on the outcome-based nature of the principle. This can be highlighted in the prescription of efficient, flexible, or effective policies. These statements focus on the characteristics of the outcome of the policies. A final, non-trivial, characteristic of the utilitarian distributive justice principle is evaluating distributions based on a costs and benefit analysis. These types of evaluations perform an aggregation of both factors, ultimately evaluating policies solely on the balanced outcome.

2.2.3 | Prioritarian

The prioritarian principle favours distributions that prioritize increasing the benefit of the worse off. Like the utilitarian principle, it is based on the outcome of the distribution, but it puts more weight on increasing the welfare of the people who are less well off. It is based on the idea that transferring resources from better off to worse off is morally desirable. The distinction between better off and worse off is made based on the lifetime level of wellbeing that can be achieved (Sinnott-Armstrong, 2023).

The principle can favour similar distributions as egalitarian, both taking the circumstances of individuals into account. However, the prioritarian principle would favour distributions of money to the worse off from the argument that it is solely the just to help het worst off, not that it would reduce inequalities. The principle is frequently found in climate policymaking, especially in the redistribution of money. In this research, redistribution of money from rich to poor seen as a prioritarian distribution. The same goes for distributions in which support is provided to the worse off.

2.2.4 | Sufficientarian

Sufficientarian principle prescribes that everyone should have a minimum threshold of resources or capabilities, ensuring that no one falls below a certain level of well-being. In contrast to egalitarianism, it doesn't prioritize eliminating all inequalities or maximizes welfare like utilitarianism but focuses on ensuring that everyone reaches this minimum standard, deeming this point as just (Arneson, 2023).

Huseby (2010) notes a positive and a negative perspective on this threshold:

1. Positive: There exists a level of advantage that is crucial for individuals to achieve. It's morally significant for people to have lives of some minimum quality.
2. Negative: Once individuals find themselves above this threshold, no further distributive justice concerns arise. The difference between individuals only slightly above the threshold and others far above is not considered and the principle no longer takes them into account and there is no reason for further distribution.

Determining the threshold itself can be challenging and is up for discussion. The level can differ per person or per context, for example in determining a specific amount of money that is needed to reach an acceptable level of welfare (Robeyns, 2017). Sufficientarian distributions are found in policies prescribing the need for redistribution to lift everyone above the poverty line. This clear indication of a certain threshold makes a policy or a distribution sufficientarian.

Where the positive sufficientarian perspective is concerned with having as many people reach the level of sufficiency, the negative perspective evokes questions on whether the upper distribution can be seen as just. The concept of **Limitarianism**, as described by Robeyns (2017), places moral limits on personal wealth for the benefit of society. She states that “no one should hold surplus money, which is defined as the money one has over and above what one needs for a fully flourishing life”. She argues that it would be fairer and more efficient for the wealthiest individuals to fund climate actions, as their surplus wealth does not contribute to their well-being but could be invested in addressing societal challenges like climate change. It essentially describes a limit to the top end of the wealth distribution as it is morally impermissible to be so rich.

2.2.5 | Libertarian

The libertarian principle of distributive justice is not concerned with outcomes or inequalities, contrary to the other identified principles. It is essentially not a preference for a distribution, but a disapproval of anything that restricts individual freedom, thus a disapproval of any forced redistribution. Davidson (2021) describes the libertarian principle of distributive justice to be based on the duty not to harm the bodily integrity and personal property of others whilst there are no obligations to help others that are worst off. It is seen as a controversial in the context of distributive justice, focussing on endorsing a free-market economy and voluntary cooperation (Van der Vossen and Christmas, 2023). Forced redistribution of wealth is seen as unjust and violating individual rights. In the context of climate policies, this principle is linked to the idea of ‘first come, first serve’ and that countries’ responsibilities to bear mitigation, adaptation, and damage costs themselves.

This principle is relevant in the context of this research as evocations can imply hesitance to any type of new climate policy. In climate negotiations, this principle can cause fundamental disputes on the moral rightness of any government intervention. It sheds light on why countries are motivated to take action and their readiness to engage with global policies.

2.3 | Distributive Justice in Global Climate Policy

The fundamental distinctions in principles of distributive justice highlight the breadth of policy discussions surrounding this concept. This breadth is also seen in policy proposals, as outlined in literature reviews of climate research and climate policymaking by Okekere (2010) and Cairney et al. (2023). They highlight that various approaches have been used to account for distributive justice in climate policies. It is concluded that there is an understanding of the need to account for the unfair consequences of climate change. However, the policies contain various approaches to allocating benefits and burdens and seldom specify the details needed for implementation.

Policy proposals are based on fairness or equity principles, essentially translating the broad distributive justice principles into rules that should be followed by everyone (Cairney, 2023). They are mostly focused on specific units of the distribution, for example prescribing equal per capita entitlements to emissions. In some cases, principles are found in the foundations of agreements, implying a widely shared ethical standard. Within the UNFCCC and global climate policymaking, there is a recognition of the Common but Differentiated Responsibilities and Respective Capabilities (CBDR) of countries. This principle does not disregard the responsibility of all countries to act, but that the level of their contribution can differ (UNFCCC, 1992). This idea is supported by the egalitarian view on distributive justice. The real-life use of this principle is less straightforwardly tied to a principle. It can be used to motivate the countries taking on self-determined voluntary action, steering away from the idea of responsibility (Okekere, 2010). In this context, CBDR is used for libertarian goal. The distributive justice principles can also be found at the foundation of policy mechanisms enforced by

the UNFCCC. An example is the Green Climate Fund, which is developed with the aim to help the poorest countries take mitigation and adaptation measures. Table 1 presents additional examples on equity principles, policy mechanisms and their corresponding distributive justice principle.

The principles of UNFCCC suggest an egalitarian foundation, emphasizing and recognizing equity. However, there is a gap between what is proposed and how justice is implemented in practice. The attention to climate change itself is higher than the attention to climate justice. This results in vague and non-commitment policy proposals, with a focus on technocratic and market-based solutions and individual responsibility (Cairney, 2023). The resulting actions and policies following this more libertarian principle are deemed to provide only limited transformative power to reduce global inequalities.

Table 1 Distributive Justice in the UNFCCC - Overview of equity principles and policy mechanisms with appointed distributive justice principles.

Equality principle / policy mechanism	Description	Distributive Justice Principle
Equal per Capita rights	"Each *person* or *country* has an equal share of allowable emissions or the same 'entitlement and burden". (Okekere, 2010)	Egalitarian
Polluter pays principle	"Each country is responsible for its own mitigation costs" (Davidson, 2021) Advocating for no interference from other countries and obligation to help the ones who are not able to bear these costs by themselves.	Libertarian
Ability to pay principle	"Individuals or countries have a responsibly to bear a larger share of the costs of climate policy the wealthier they are, irrespective of the extent to which they have contributed to climate change" (Davidson, 2021) Notion of every country needing to take responsibility, whilst acknowledging differences.	Egalitarian
Basic needs	Allocate minimum emission rights necessary for survival.	Sufficientarian
Market Share Principle	Create tradable permits for emission to achieve lowest net world cost for abatement by the free market (Okekere, 2010)	Libertarian
Mutual advantage	'Allocate benefits and burden' to ensure a 'positive net benefit for all' (Okekere, 2010)	Utilitarian
Green Climate Fund	Financial mechanism of the UNFCCC, part of the Paris Agreement, established to (financially) support developing countries in realizing their climate ambitions (GCF, n.d.).	Prioritarian
Common but Differentiated Responsibilities	"The Parties should protect the climate system for the benefit of present and future generations of humankind, on the basis of equity and in accordance with their common but differentiated responsibilities and respective capabilities." (UNFCCC, 1992) Developed countries[4] bear a greater responsibility and should take the lead in the taking action. Developing countries[5] should be supported in their climate change activities by developed countries with financial assistance (UNFCCC, 1992).	Egalitarian

[4] Developed countries indicate countries that are industrialized, with high standards of living and strong economic growth. The classification is made based on factors like GDP. Within the UNFCCC they are referred to as Annex I countries and are expected to lead the way in the policymaking process (UN, 2014).

[5] Developing countries have a relatively low standard of living and have not yet achieved a significant degree of industrialization. They area also known as Non-Annex I parties, recognized by the UNFCCC as being especially vulnerable to the impacts of climate change.

2.4 | Main findings

Distributive justice covers three dimensions: scope, unit, and shape. These elements can be seen as notions of the practical implementation of distributive justice. What elements seen as important in these dimensions can be motivated by principles of distributive justice identified for these three dimensions. Five key principles have been identified and defined in the context of climate policy. Egalitarianism advocates for reducing inequalities and ensuring fairness in the distribution of resources, opportunities, and responsibilities. Utilitarianism prioritizes maximizing overall welfare or benefits, disregarding individual, or subgroup differences. Prioritarianism emphasizes increasing the welfare of the worst off, aiming to uplift those in the most disadvantaged positions. Sufficientarianism focuses on ensuring that everyone reaches a minimum threshold of well-being, while libertarianism emphasizes individual freedom and opposes forced redistribution. Within climate policy proposals various equality principles and policy mechanisms are found that can be tied to distributive justice principles.

3 | Methodology

As highlighted, there is a need to understand the use of distributive justice principles in climate policymaking. Distributions are mentioned in various contexts and the principles function as moral justifications or policy choices. This chapter presents the workflow applied to identify the use of distributive justice principles in climate discourse. First an overview of the research workflow is presented, followed by a discussion of the method used in the two research parts.

3.1 | General Approach

The multilayered definitions of distributive justice and its principles present implications for the identification of its use in texts. The analysis should take the complex and nuanced nature of the concepts into account as well as account for the contextual understanding of distributive justice in climate policymaking. The overarching method used in this research to capture these complexities is qualitative coding. This method identifies patterns and themes by using a predetermined codebook to label textual data into a fixed set of codes, listing all labels (i.e. codes) with definitions and examples (Xiao et al., 2023). The codebook is used to classify text or textual elements to specific labels. This approach presents the underlying assumptions of the analysis and ensures consistency when annotating large amounts of data by one or multiple coders (MacQueen et al., 1998). The resulting coded data can be used to derive theories and increased understanding of the texts themselves (Hsieh et al., 2005).

Qualitative coding is a time-consuming process, including both the formulation of a codebook, iteratively improving it, and eventually applying the codebook to evaluate all data, preferably by multiple annotators. Additionally, human annotators can be subject to changing perspectives during the annotation process, which can lead to labelled text data that contains inconsistencies and errors (Pangakis et al., 2023). To assist in the coding process, various AI-based tools that use natural language processing and machine learning algorithms have been developed. These methods can be supervised, e.g. using large datasets to perform logistic regression, or unsupervised, e.g. topic models, to help discover themes found in texts (Xiao et al., 2023, Bagozzi, 2015). These task-specific tools have their limitations, either calling for large high-quality datasets or difficulties to use model outputs for more complex research questions.

LLMs address both the challenges of qualitative analysis as well as overcoming the limitations of the previously described AI-based tools. LLMs have the capacity to learn representations of words and patterns in language. They can be instructed to perform specific tasks with textual prompts, including specific instructions and examples, like codebooks and are able to generate textual answers (Xiao et al., 2023). Numerous studies present varying results and recommendations on the application of LLMs for text annotation tasks or potentially replacing human annotators. Some indicate that LLMs, specifically OpenAI's GPT, show similar accuracy to human annotators in annotation tasks (Ding et al. 2023; Gilardi et al. 2023; Törnberg, 2023a). Savelka et al. (2023) explicitly highlight the capability of GPT-4 to analyse text that require highly specialized domain expertise, in their case legal analysis. Others, like Pangakis et al. (2023), highlight that the performance of LLMs varies across annotation tasks, evaluating over 27 different tasks with GPT-4. They point to the to the quality of the prompt, textual 'quirks' indicating unconventional textual behaviour, or conceptual difficulty of the task, as factors influencing the performance.

3.1.1 | Model selection

Most LLMs are based on a transformer structure, breaking down words into tokens. They predict words based on the context, where the contextual relationship is captured with a self-attention mechanism, essentially using statistics decide what words to generate (Vaswani et al., 2017). The context can be taken into account unidirectional, only accounting for previous words, or bi-directional taking both previous and next words into account.

There are several types of LLMs available, differing fundamentally in terms of accessibility and transparency. Closed source models, like OpenAI's GPTs, are not publicly owned and model architecture is not accessible. It is known that OpenAI's models are trained on 45TB of raw data, they present limited transparency what is included in these datasets (Brown et al., 2020). This leads to limited accessibility for users to the model's development and training processes (Ouyang et al., 2023). Additionally, as OpenAI is a commercial institution, it charges a fee to access the model.

Models like Google's BERT are open source and are freely available to the public. They can be modified easily, and underlying architecture and dataset are mostly publicly available. BERT was trained on 3.5 TB of raw data, consisting of Wikipedia entries and Google Books (Delvin et al., 2018). It is open to community contributions, especially as the they can be trained for specific tasks by means of fine-tuning. An example of a modification is ClimateBERT, a version of Google's BERT. ClimateBERT is a version of Google's BERT, pretrained by non-google affiliated researchers on climate-related texts and can be used to classify and fact-check climate related texts (Webersinke et al. 2022).

Although these open-source models are used increasingly for several types of tasks, their performance has not been proven like OpenAI's GPTs for textual annotation tasks (e.g. Savelka et al.,2023, Gilardi et al. 2023). This, combined with increasingly widespread use of GPTs in qualitative research, has led to the choice to evaluate OpenAI's GPT performance for the distributive justice identification task. This research uses GPT-4o, OpenAI's most recent flagship model (OpenAI, 2024).

3.1.2 | Workflow

This research follows a workflow for LLM augmented text annotation as presented by Pangakis et al. (2023) which was created to ensure human-in-the-loop updates of LLM prompts as well as enable human validation of its labels. The workflow, as visualised in Figure 1, consist of two parts. In part A, codebooks are created deductively. The theoretical foundation for this codebook is outlined in Chapter 2, defining the principles of distributive justice. The resulting codebooks are used to annotate a subset of speeches and updated inductively with observations made in this process. The codebooks and annotated dataset are used in part B, employing GPT-4o with the same annotation task and evaluating its performance. The performance of the model is evaluated based on accuracy, precision, recall, and F1. All data created and used in this research is available on Github.

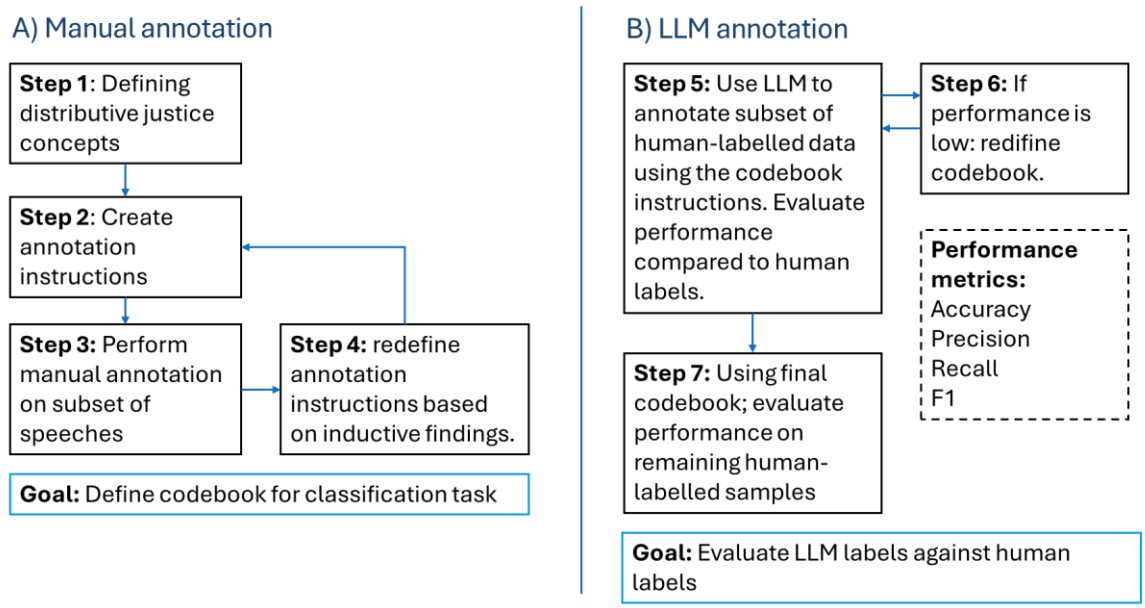

Figure 1 Research workflow, adapted from Pangakis et al. (2023). A distinction is made between two research components, with the first focussing on the development of a codebook and human-annotated subset of data and the second focussing on the evaluation of LLM performance.

3.2 | Manual annotation method

A corpus of 1063 speeches is created with documents scraped from the public UNFCCC database covering COP19 to COP28. 151 of 198 parties made at least one of their speeches available in English, resulting in a total of 744 documents. On average, there are seventy-four speeches per COP available with an average length of 4079 characters. More information on the HLS corpus and speech availability is found in Appendix A.

The manual annotation is conducted on a subset of HLS speeches. The subset is made by randomly selecting five speeches per COP, resulting in a subset of fifty speeches covering 6% of all English HLS documents and 50 different countries. A visualisation is presented in Figure 2, Table 2 presents an overview of speeches per COP. During the manual annotation, full speeches are evaluated, but the annotation is sentence based, meaning that each sentence will be labelled individually, resulting in 1543 labelled sentences. Limited time forces representatives to cramp the countries focal points in short text, leading to high topic density in the HLS (Bagozzi, 2015). With high topic density, it is expected that most sentences discuss a different subject and potentially a different distribution. However, as sentences can refer to topics of previous or next sentences, the previous and next two sentences are seen as context during the annotation.

The manual annotation is conducted in Excel, with each row presenting a sentence and annotation columns containing the annotation categories and a short (max. 30 word) argumentation. The first step of annotation consists of reading all sentences and determining their relevance. Sentences are deemed relevant if they present a normative statement with indication of moral judgement or preference. Next the full text is evaluated again, annotating the relevant sentences for the motivational distributive justice principle and three additional categories that can be used to evaluate the context in which principles are identified. The first category is Topic, identifying the main topic of the sentence, allowing to evaluate if certain principles are found more often when a specific topic is discussed. The other two categories refer to the distributional dimensions Unit and Shape. These provide information on how the distributive justice principle should be implemented in practice and to see if these correlate with distributive justice principles. After full evaluation of all speeches, inductively determined labels and definitions are summarized into a codebook.

For each annotation category a codebook is designed that follows the structure of [Code: / Description: /], following Xiao et. al. (2023). The considered judgement that is applied when confronted with a conceptual gap between theoretic and real-life use of the principles, is illustrated with examples. The created ground truth dataset with labelled sentences is evaluated with exploratory data analysis, evaluating frequencies of occurrences per speech, class imbalances and potential correlations between labels of multiple categories.

Table 2 Manual annotation - Overview of randomly selected HLS speeches for manual annotation. In total, 11 of the speeches are presented by Annex-1 countries, indicated in blue. This ration is similar to the ratio of Annex-1/non-Annex-1 in the UNFCCC (21%).

***** *Tonga - Pacific Small Island Developing States: The HLS speech presented by Tonga included an additional speech on behalf of PSIDS, one of the negotiation groups at COP. It was chosen to include this speech separately in the subset*.

COP19	COP20	COP21	COP22	COP23
Japan	Holy Sea	Afghanistan	Belize	Croatia
Namibia	Kenya	Nepal	Czechia	Ireland
USA	Micronesia	EU	Israel	Netherlands
Timor Leste	Tonga *Tonga PSIDS**	Bosnia and Herzegovina	Lao	Suriname
Sierra Leone	Republic of Korea (*South Korea)*	Trinidad and Tobago	Thailand	Vanuatu

COP24	COP25	COP26	COP27	COP28
Indonesia	Cambodia	Barbados	Australia	Belgium
Lesotho	Malaysia	Grenada	Greece	Gambia
Macedonia	Mauritius	Philippines	Kazakhstan	Jamaica
Malawi	Serbia	Russia	Serbia	Jordan
Rwanda	Uganda	Slovakia	South Sudan	New Zealand

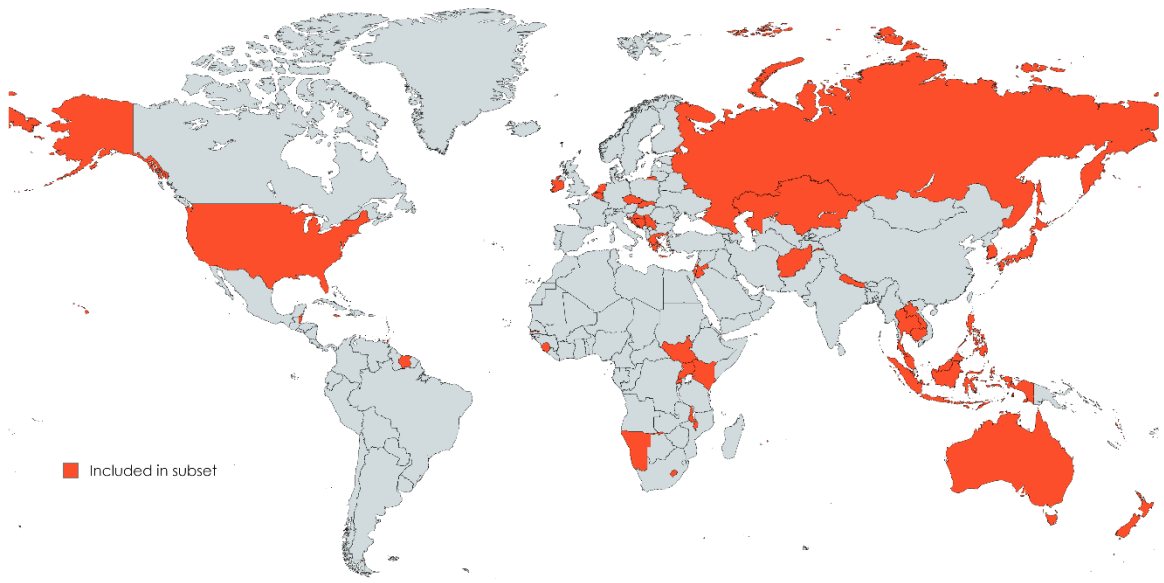

Figure 2 Visualisation of the randomly selected HLS speech subset. It is important to note that a very limited number of parties in Middle and South America, North Africa, and the Middle East have made their speeches available in English.

3.3 | LLM annotation method

The derived codebooks and ground truth dataset are used during LLM annotation, where the codebooks are converted into prompts. The GPT-4o model is employed with an API, in a workflow adapted from Pangakis et al. (2023). Following instructions provided in a prompt, batches of 20 sentences are annotated for the categories, tasking the model with selecting one of the pre-defined labels that apply to a sentence. Batches are used because model requests are limited by tokenlimits, limiting the number of characters that the combined prompt and model completion (answer generated by the model) can contain. Requests can include up to 128.000 tokens (+- 500.000 characters), this is not sufficient to evaluate all sentences in one request. The batch size of 20 was selected arbitrarily.

3.3.1 | Reproducibility

To ensure reproducibility of this study, model outcomes must be deterministic. This is challenging due to the non-deterministic nature of GPT models, presenting differing outcomes for the same task (Ouyang et. al., 2023). However, several steps are taken to ensure, to the extent possible, that annotations presented by the model are of minimal variability:

1. Outcomes in specified format:
   The option to vary outcomes is limited by the nature of the annotation task itself. A set of prescribed labels, as derived during manual annotation is presented to the LLM, limiting the output space. All primary evaluations are performed with string-based labels. The use of binary labels is evaluated for a single annotation task.

2. Temperature = 0 – *default: 1*
   The sampling temperature parameter determines the randomness (also named creativity) of the text generated by the model by weighing the likelihood of outcomes. A higher temperature results in a more diverse output. A temperature of 0 is selected to minimize this randomness by selecting the most likely output. By setting the temperature to 0, determinism of the outcomes is improved compared to the default setting of 1, but non-determinism is not completely avoided (Ouyang et al., 2023). All primary evaluations are performed with a temperature of 0. The influence of using a non-zero temperature is evaluated for a single annotation task.
3. Top_p = 1 – *default: 1*
   This alternative sampling parameter influences the variability of outcomes with nucleus sampling. It is recommended not to alter this parameter when altering temperature (OpenAI, n.d.). In this research it is set to 1.
4. Seed-based evaluation – [3644,3441, 280, 5991, 7917]
   The seed controls the reproducibility of the request. If specified, the model makes a *best effort* to sample deterministically (Anadkat, 2023). Presenting the same seed and parameters for the same task presents leads to mostly consistent outputs. This research evaluates model performance for five randomly selected seeds. Each seed indicates a specific instance of the model that is called upon by the API for each batch of sentences.
5. Post-call filtering on system_fingerprint
   The system_fingerprint is the backend configuration used during the model run, which can vary with each API call and cannot be controlled by the modeler. Evaluations made for the same seed with the same system_fingerprint should ensure mostly deterministic outcomes (Anadkat, 2023). They are part of the API response and can only be evaluated after the API call is made. As the data is evaluated in batches, each model instance is called upon multiple times until all sentences are annotated. Post-call filtering is applied to omit results generated with a different backend configuration. If the fingerprint of a batch annotation does not match the fingerprint used in the majority of API calls, the annotations made for this batch are omitted. The number of omitted sentences is noted and only sentences that are annotated by the GPT for all five seeds are evaluated.
6. Marginal differences
   For the same seed, fingerprint and prompt, the model should generate the same results. However, in rare cases, the non-deterministic nature of the model can lead to marginal differences, even though the same seed, fingerprint and model parameters are used (Ouyang et al., 2023). Occurrences of these marginal differences are evaluated in secondary testing by performing an annotation task three times. The limited occurrence of these instances, combined with temporal and financial limitations, have led to the decision to perform all primary evaluations with a single iteration.

3.3.2 | Performance metrics

The model performs a multi-class classification task, where each sentence is labelled as a single principle, that can be compared to a 'ground truth' dataset. The performance of the model is evaluated based on multiple evaluation metrics: accuracy, precision, recall, and F1 (Rainio et al., 2024). This research compares the model performance to thresholds used in other research evaluating the performance of models in specific classification tasks (e.g. Pangakis et al., 2023; Ding et al., 2023, Gilardi et al., 2023; Balagoopalan et al., 2023). Metrics below 0.5 are deemed unsatisfactory.

$$(1)\ Accuracy = \frac{Correct\ predictions}{All\ predictions}$$

Accuracy evaluates the number of correct predictions and presents a general estimation of performance. This metric makes no distinction between different classes, which can lead to a distorted picture of model performance as majority classes are easier to predict. There is no distinction between the different classes, which can be problematic in case of class imbalances.

$$(2)\ Precision_{label} = \frac{Correctly\ predicted_{label}}{Total\ predicted_{label}}$$

Precision measures the ratio of correctly predicted instances of a label (i.e. class) to the total number of that label that is predicted. It indicates how often the model is correct when it predicts a specific label. Precision is calculated individually for each label and a macro average over all labels is determined. This average does not account for class imbalances between labels.

$$(3)\ Recall_{label} = \frac{Correctly\ predicted_{label}}{Total\ Ground\ Truth_{label}}$$

Recall measures the ratio of correctly predicted labels to the to the ground truth labels of that class. It indicates how much of the ground truth labels are captured by the model. Like precision, recall is evaluated individually and averaged over all labels.

$$(4)\ F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Precision and recall often present a trade-off. In this research, it is both important that annotations are correct. This means that both accuracy and recall are equally important. The F1 score presents the harmonic mean of these two metrics, enabling their joint evaluation per label.

3.3.3 | Experiments

The dataset created by manual annotation is used as the 'ground truth' to which the model classifications are compared. As both codebook and the data are generated in this research, the data has not been seen by the LLM before. Four primary annotation tasks are defined, for which up to four string-based task-specific prompts are formulated. These tasks are primary as they are all evaluated with the same five seeds, a single iteration and a temperature of 0.

The use of multiple task-specific prompts can be seen as version of prompt-engineering. Two prompt settings are evaluated: zero-shot and one-shot. The zero-shot prompts only indicate a label and a definition. By using the same codebook as during the manual annotation, a conceptual gap between human and model instructions is prevented, which is crucial when model performance is compared to human annotators (Pangakis et al., 2023). Only additional instructions in the prompt are provided to constrain the output space, defining the structure of the response. The one-shot setting follows the same structure but includes one example for each label. Tasks are also evaluated with two context-related instructions. The first does not specify anything about the evaluation of previous or next sentences. The second specifies that the previous and next two sentences are to be taken into consideration when a label is selected. An overview of all primary annotation tasks is presented in table 5. All prompts are available in Appendix F.

Table 3 LLM annotation - Primary experiment overview - All experiments are performed with the same 5 seeds, 1 iteration and a temperature of 0.

			Prompt			
			Zero-shot		One-shot	
Task	**Class**	**Dataset**	-	Context	-	Context
B1	Relevance	Full	B1.0	B1.0.1	B1.1	B1.1.1
B2	Principle	Relevant sentences	B2.0		B2.1	
B3	Principle	Full	B3.0	B3.0.1	B3.1	B3.1.1
B4	Principle, Topic, Unit, Shape	Full, Relevant sentences	B4.0			

Task B1, B2, and B3 are multi-class classification tasks, where each instance is assigned a single label (i.e. class). Task B1 is the pre-processing task. This task is performed on the full train dataset. Task B2 is performed on a subset of the training dataset that includes only the relevant sentences. It

classifies the relevant sentences for the distributive justice principle used. As only relevant sentences are annotated, it is not possible to evaluate the influence of contextual sentences. In task B3, principle classification is performed on the full training dataset, labelling relevant sentences for principles and identifying non-relevant sentences as such.

The LLM performance on classifying distributive justice principles and the three characteristics is evaluated in task B4, a multi-label classification task. It is performed on the full training dataset, leading to the inclusion of the additional 'not evaluated' label for each of the four categories, and separately on the subset of relevant sentences. This task is the most elaborate and is only evaluated in a zero-shot setting.

Four secondary annotation tasks are formulated to evaluate the effect of model settings and one-hot coded (binary) labelling for each label. All tasks evaluated the full dataset for relevance, using prompt B1.0 and vary in the number of iterations and temperature. For the numerical annotation, a numerical prompt (A1.0) is created. Table 6 presents an overview of all secondary experiments.

Table 4 LLM annotation - Secondary experiment overview

Task	Prompt	Iterations	Temp.
S1	B1.0	1	0.6
S2	B1.0	3	0
S3	B1.0	3	0.6
S4	A1.0	1	0

4 | Part A: Manual annotation

The findings of the conceptualisation and insights gained during annotation are used to create a codebook for each of the categories. This chapter describes the insights gained during general manual evaluation, the codebooks created, and explores the created ground truth dataset.

4.1 | Codebooks

On average, full annotation of a single speech took 20 minutes. Most speeches follow a similar pattern: expressions of gratitude to the host country, notions of consequences of climate change currently experienced in the speaker's country, the current state of climate policies in the speaker's country, followed by expectations on the outcome of the COP and references to the need for urgency and cooperation. Countries emphasize different climate-related consequences and national policies. Examples are Suriname highlighting their role in forest conservatory and Malaysia focussing on renewable energy.

As the speeches are a version of spoken word, they present room for semantic interpretation. Upon evaluating speeches, tonal and narrative related differences can be found. The USA and Laos focus on highlighting personal milestones, where Island states like Vanuatu and Barbados highlight the consequences of climate change for both the world and themselves. Developing countries present themselves frequently in alliance with other developing countries. Developed countries seem to be more focused on personal efforts. An additional difference can be found in a change in narrative before and after the Paris Agreement (COP21), where pre-Paris speeches present more policy expectations and post-Paris focuses on implementation.

All these elements add to the need to evaluate full sentences and contextual meaning, where there is room for interpretation and steering away from a word-count-based analysis. This can also be tied to the gap between theory and practice, meaning that theoretical knowledge and created definitions of principles could not straightforwardly be applied to a sentence. Considered judgement is applied to these sentences, founded in the theory, but always subject to potential interpretational differences between multiple annotators. This does not only count for the defined principles, but also in the determination of relevance, topic, unit, and shape.

The next sections address each of the categories and uses example sentences to illustrate the considered judgements made. The broad range of topics, meaning, and interpretational possibilities make it unfeasible to cover every edge case. The created codebooks contain general definitions of the labels, where room for interpretation is maintained. All full codebooks are available in Appendix C-E.

4.1.1 | Pre-processing

Not every sentence of a HLS speech is relevant to evaluate for distributive justice characteristics. Table 5 presents eight different example sentences of various relevance.

Speeches contain formalities, thanking host-countries or expressing condolences with respect to (natural) tragedies [1]. Only relevant sentences can be labelled for elements of distributive justice. However, defining relevance is challenging due to the broad range of concepts that is discussed and the textual ambiguities that arise in written texts. Notions of distributions of resources, actions, or policy measures are made in different contexts, not always presenting indications of motivational principles, or knowingly leaving room for interpretation [2,3].

With the aim of identifying distributive justice principles, relevance is defined as sentences presenting normative statements. Normative statements involve value judgements or prescribe actions based on norms and values. In contrast to descriptive statements that objectively describe reality, they cannot be checked to be true, expressing opinions or beliefs. The values evoked in normative statements,

especially in relation to new policies, can be related to the distributional principles that are the motivation behind these normative claims [4]. Not relevant sentences are, apart from formalities, descriptive, highlighting facts about current policies or the state of the world [5]. The distinction between descriptive and normative is not binary, especially in texts where (vocal) emphasis can be applied to words in such a way that the meaning of the word changes (Van de Poel & Royakkers, 2023). This leads to interpretation bias, written text presenting room for multiple interpretations, especially when limited normative context is presented.

To account for ambiguous sentences a third relevance label is introduced. This “statement of intent” label presents a midway category for sentences that can be deemed both normative and descriptive, often lacking clear indications of specific normative motivations [6,7]. The label name is originally derived from sentences that present new policies that parties are intending to implement. These intentions can be interpreted as presenting some indication of what countries find important, as they indicate that actions will be taken, but the normative motivations are mostly open to interpretation of the reader. A final distinctive type of sentence are rhetoric questions, often used in speeches to play upon the feelings of a listener [8]. Although these sentences can be interpreted as implicitly displaying normative judgement, or at least trying to evoke some normative judgement with a listener, their main function is to create engagement. This element, combined with their lack of direct expressions of normative preferences, leads to defining these sentences as not relevant.

Table 5 Manual annotation - Preprocessing for relevance: example sentences. Additional examples with label argumentations can be found in Appendix B2.

HLS Speech	Sentence	Notes
[1] COP21 Nepal	At the outset, I would like to express our deepest condolences on the recent tragic incidents in Paris and convey solidarity with the people and Government of France.	Expression of formalities, in this case condolences to the government of France. Labelled as: Not relevant
[2] COP24 Indonesia	We could not accept the use of environmental issues including climate to be a means of disguise trade discrimination against developing countries.	Does not explicitly present a distribution of resources but does advocates against discrimination, indicating the principle of egalitarian as motivation. Labelled as: Relevant
[3] COP27 Australia	We have a moral imperative and driving need for our institutions to work with countries across the developed and developing world.	An implicit reference to a distribution, in the sense of the need for cooperation. Labelling for relevance based on the explicit mentioning of resources would have excluded this sentence. Labelled as: Relevant
[4] COP19 Namibia	It should not only enable us to discuss global environmental challenges our world faces, but more importantly, it should pave way to key decisions that would put climate change action at the fore-front of all developmental processes and giving hope to our future generations.	Normative statement prescribing the need to make key decisions with regards to climate change, motivated by giving hope to future generations. This motivation can be seen as Utilitarian. Labelled as: Relevant
[5] COP20 Kenya	As part of the roadmap agreed in Durban, Doha and Warsaw, Lima must produce the elements of a draft negotiation text for the post -2015 Agreement in accordance with the objective, principles and provisions of the Convention.	Describes the characteristics a new policy should adhere to. A normative element can be found in the reference to the foundations of the UNFCCC, which is inherently egalitarian. Labelled as: Statement of intent
[6] COP19 Timor Leste	In addition, all parties need to reaffirm their efforts in addressing the ratification processes for the second commitment period of the Kyoto Protocol in order to come into force as soon as possible.	Prescribes other parties to ratify commitments made. This prescriptive element can be expressed with normative emphasis or without. Labelled as: Statement of intent

[7] COP19 Sierra Leone	Typhoon Haiyan has once again gravely reminded us of the looming disaster and the urgent imperative to put aside the endless rhetoric and act NOW.	Emphasis on the urgency, indicating that this is motivated by norms and values and not a description of expectations. Labelled as: Relevant
[8] COP26 Barbados	Do so me leaders believe they can survive and thrive on their own?	Rhetoric question, implicitly prescribing the need for cooperation. Labelled as: Not relevant

4.1.2 | Distributive Justice Principles

Sentences are labelled for one of five principles: Egalitarian, Utilitarian, Prioritarian, Libertarian, or Sufficientarian. Determination of these principles is based on the understanding of the normative context in which the statement is presented. The labels classify the norms and values on which normative statement are founded. The labelling is subject to some interpretation bias as it is not always clear what motivation is used. In these cases, sentences are labelled as General normative statements. At times, foundational elements of the UNFCCC are highlighted. The identified principles behind these elements, as mentioned in Chapter 2, are taken into account during labelling. Table 6 presents four examples of labelled sentences.

Table 6 Manual annotation - Labelling for Principles: example sentences. More examples and argumentations can be found in Appendix B.

HLS Speech	Sentence	Notes
[1] COP19 Namibia	In addition, major developing countries must also reduce their emissions through National Appropriate Mitigation Actions (NAMAs) and other form of mitigation actions.	Prescribing and calling on the need of developing countries to reduce their emissions. This could be from an egalitarian perspective, with the idea that everyone should contribute, but also from a utilitarian foundation, calling on the need for implementation measures for the benefit of all. Labelled as: General Normative Statement
[2] COP20 Holy Sea	Let us work together for the common good rather than point at each other and pass responsibility to others.	Prescribing the need for cooperation. The specific highlight of “the common good” refers to the utilitarian motivation of acting in the benefit of all. Labelled as: Utilitarian
[3] COP25 Malaysia	We would like to recall that the convention obliges developed country Parties to lead mitigation actions.	Directly pointing the foundation of the convention and the principle of CBDR. Labelled as: Egalitarian
[4] COP23 Vanuatu	The global response to climate change must put fairness and equity at the heart of its work, and to keep the needs and aspirations of the world’s most vulnerable countries in its line of sight.	Highlighting the normative ideas of fairness, equity and helping the most vulnerable. This last element could point to a prioritarian motivation, but its combination with fairness and equity makes reducing inequalities the main motivation. Labelled as: Egalitarian
[4] COP20 Tonga	We must all work collectively, with a sense of urgency and purpose, to address these challenges and to support all island countries to become more resilient.	Highlighting the need for cooperation, with the aim to help the small island countries (that are most vulnerable to climate change). This makes the motivation prioritarian and not motivated by reducing inequalities. Labelled as: Prioritarian

Some sentences can be interpreted as indicational of multiple principles. An example is sentence 3, where the principle of CBDR is seen as an indication of the egalitarian principle. However, it can be argued that this principle is used in a prioritarian sense, where Malaysia could be motivated by the idea that CBDR should be applied to support the less fortunate.

Full sentences and semantic contexts are used to determine the distributive justice principle that is highlighted in a sentence. To summarize the essence of the various sentences, wordclouds are used. They present the 30 most used words, enabling comparison between principles and determining distinctive terms. Figures 3-7 present the wordcloud for each of the principles, with word size proportional to frequency.

The wordclouds highlight that word use can be tied to the principles, for example with support, finance, vulnerable more often named in sentences presenting a prioritarian principle. The themes determined in the theoretical concepts are also found in real life speeches. Utilitarian highlights the future and mentions the word global more often than other principles. Egalitarian sentences also note this global element, but also use words like fair, including, together, and agreements. Sufficientarian sentences mentions development, like the prioritarian principle, but also focusses on words like necessity and mentions the word behind as an indication that there is a line that everyone should be able to reach. Libertarian sentences mention economy, ambition, work and businesses, steering away from themes related to support, responsibility, and the future.

Figure 3 Egalitarian – Wordcloud of 30 most frequently used words excluding stop words.

Figure 4 Prioritarian - Wordcloud of 30 most frequently used words excluding stop words.

Figure 5 Utilitarian - Wordcloud of 30 most frequently used words excluding stop words.

Figure 6 Sufficientarian - Wordcloud of 30 most frequently used words excluding stop words.

Figure 7 Libertarian - Wordcloud of 30 most frequently used words excluding stop words.

4.1.3 | Categories

The aim of the annotation process is to label for principle, but to evaluate the context in which distributive justice principles are presented, relevant sentences are annotated for three additional categories: Topic, Unit, and Shape. The Topic contains the context in which a principle is indicated. The dimensions Unit and Shape provide information on how distributive justice is implemented in practice, covering what is being distributed, and how the distribution occurs. The number of categories has been limited to shorten the annotation task and not all characteristics were clearly identifiable in the texts. This can be explained by the diplomatic nature of the speeches and the context in which they are presented. The initial annotation of the speeches also included the evaluation of scope and time, but these categories have been omitted due to lack of diverse outcomes. Scope of distributive justice would mostly refer to global distributions, as it is a global convention, and timely references do not include more than a vague mention of “future generations”.

Each sentence is appointed one label per category. Each category and the identified labels are briefly defined, with references to examples presented in Appendix D. The appendix also includes the codebooks defined for all category labels. Table 7 presents an overview of all determined labels for the three categories.

Topic

To enable quantitative analysis of the context in which distributive justice principles are presented, sentences are labelled for the main topic they discuss. The topics are derived inductively from the manually annotated data and aggregated into groups. This aggregation limits the diversity of topics available in the final dataset but is necessary as the annotation process works with a pre-determined list of labels. normative sentences refer to requirements for new UNFCCC policy or existing UNFCCC agreements and principles [F3]. Others present normative motivation as to why urgency or cooperation are required [F2, F4]. Financial mechanisms, like the Green Climate fund, are also frequently described and a reasoning is presented on what should happen with regards to financial redistributions [F1]. The final topics are adaptation and mitigation, which also occur together. These topics often refer to specific policies that should be implemented [F5]. All sentences that have a topic that does not fit in one of the aggregated categories is labelled as ‘other’ [F6].

Unit

Labelling for the units presents insights into both the context and practical implementation of principles. As the notion of a distribution can be implicit, the unit can be as vaguely described or not even indicated [F5]. When units are described in these speeches, they often refer to aggregated units like technological resources or general support. What exactly is meant by technological resources or support is often not indicated [F7]. The most quantitative unit is money, a driving forces in global climate policymaking, which is labelled as financial resources [F1]. The distribution of climate responsibility is also highlighted. It is found in the context of a responsibility to take on action, implement policies, or adhere to previous commitments [F2, F6]. The distribution of this responsibility can refer to specific countries or a more general call to all.

Shape

Like the unit of distribution, the shape cannot always be derived [e.g. F1, F2]. Classical shapes like equality, equality, and priority to the worst off can be found in some sentences, although not always highlighted explicitly [F6]. The manual annotation identified three additional shapes: proportional to contribution, proportional to commitment, and proportional to needs. The first shape can be connected to the idea of historical responsibility, indicating that countries that have emitted more in the past now have a larger responsibility to reduce emissions [F9]. Proportional to commitment refers to distributions that are based on international agreements [F8]. Especially in the context of financial mechanisms, a needs-based distribution is evoked. This type of distribution advocates for distributions that are based on the needs of parties, for example to implement technological innovations [F7].

Table 7 Manual Annotation - Label overview of Topic, Unit, Shape

TOPIC	UNIT	SHAPE
New UNFCCC Policy	Not indicated	Not indicated
UNFCCC agreements and principles	Responsibility	Equality
Urgency	Financial resources	Equity
Cooperation	Technological resources	Priority to worst off
Financial mechanism	Financial and technological resources	Needs based
Adaptation	Support	Proportional to contribution
Mitigation	Other	Proportional to commitments
Adaptation and mitigation		
Other		

4.2 | Ground truth dataset

During the manual annotation process, a dataset consisting of 1543 annotated sentences is created. The dataset is used as the 'ground truth' dataset in the next part of this research. Two versions of the ground truth dataset are created and available on Github. The first dataset presents annotations before all pre-defined labels were set. This means that the labels that are later covered under the 'other' are still available. The second dataset includes only the pre-defined categories, with full label names and a numerical representation of labels. Both datasets include short annotation argumentation for the relevant sentences. Exploratory data analysis is performed on the pre-defined ground truth dataset to evaluate some preliminary patterns and understand the characteristics of the dataset.

4.2.1 | Pre-processing

In total, 267 sentences are labelled as relevant (17%) and 277 as statement of intent. The remaining sentences, accounting for 65% of the dataset are labelled as not relevant. On average, speeches contain 5 relevant sentences. Figure 8 visualises the label counts for relevance for each of the annotated speeches. It shows that not all speeches are of equal length, but longer speeches do not necessarily lead to more relevant sentences.

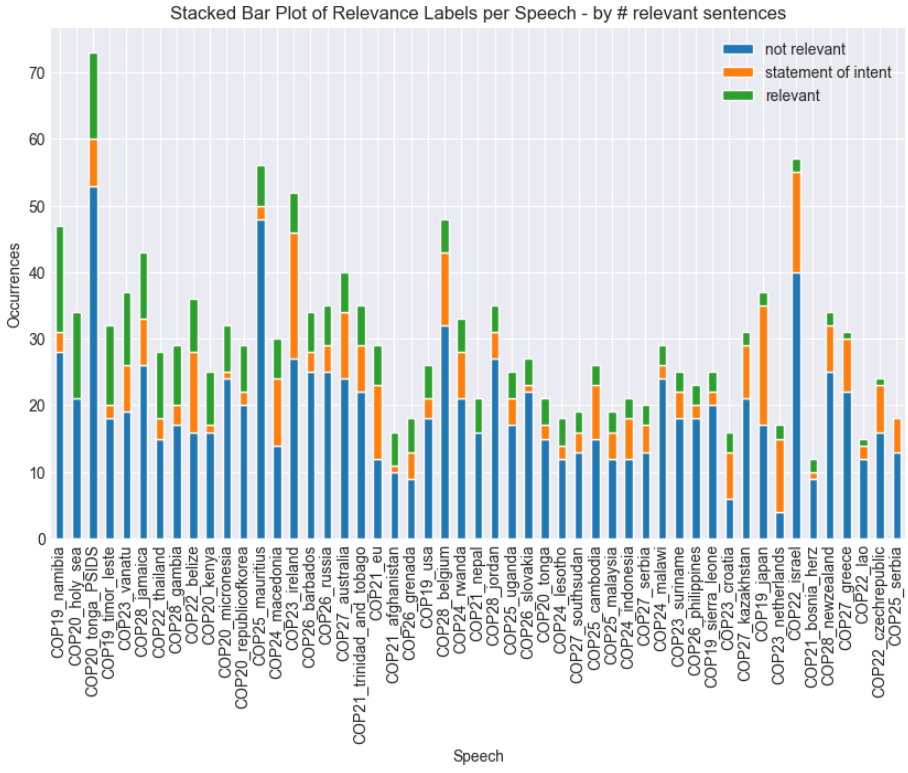

Figure 8 Manual annotation - Stacked bar plot of relevance labels per manually annotated speech.

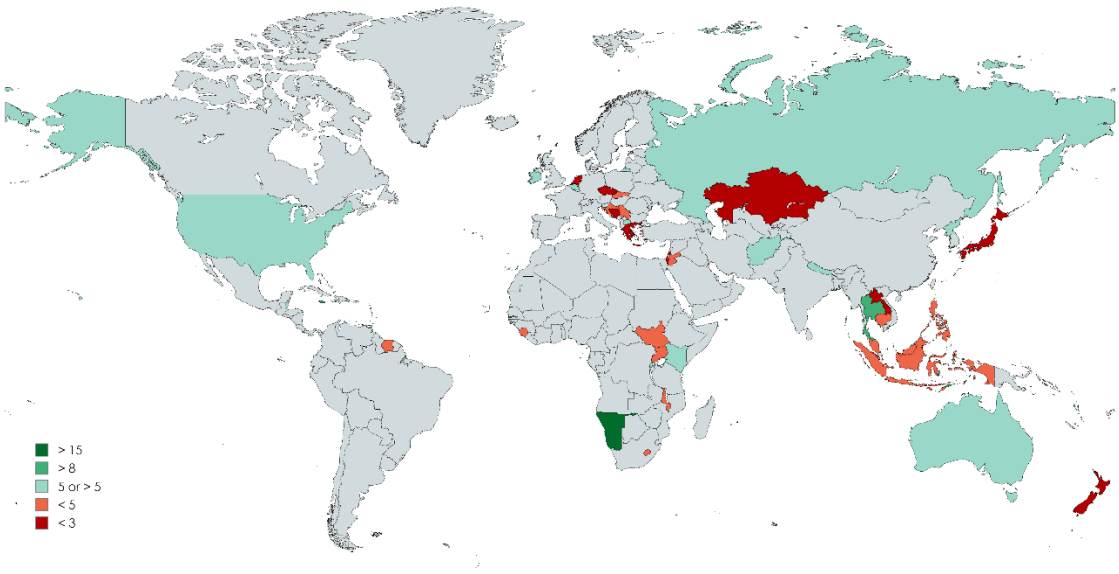

Figure 9 Manual Annotation - Number of relevant sentences per country speech.

Figure 9 visualises the number of relevant sentences per country. The speech given by Serbia during COP25 did not include any relevant sentences, outlining current practices or future policy implementations. Serbia's COP27 speech included 3 relevant sentences, still only accounting for 1.5% of the full speech. Namibia's COP19 speech included 16 relevant sentences, the largest number, mostly outlining expectations and beliefs that should be included in new policies. More than half of The Netherlands' COP23 speech included statements of intent, where Nepal and the Holy Sea did not present any of these sentences. Smaller (island) states and party groupings like PSIDS present, on average, more relevant statements than Annex I parties.

These results reveal large class imbalances with only a small portion of the speeches presenting relevant statements. Only relevant statements are annotated for categories related to distributive justice, significantly reducing the number of data points available for evaluating model performance.

4.2.2 | Distributive Justice Principles

All principles are identified within the evaluated speeches, but large class imbalances can be found. The label libertarian label is rare and only applied to four sentences in speeches from the USA, Belgium, Rwanda and The Netherlands. The Sufficientarian label is applied to 4% of the sentences. Prioritarian labels are the most occurring, with speeches highlighting this principle often more than once. As shown in Figure 10, the labels Egalitarian, Utilitarian and Prioritarian are balanced. General normative statements are less often indicated, but still in almost 20% of the sentences. This finding indicates that, even though a relevant statement is indicated, it is not always possible to identify the motivational principle that is fundamental to this statement.

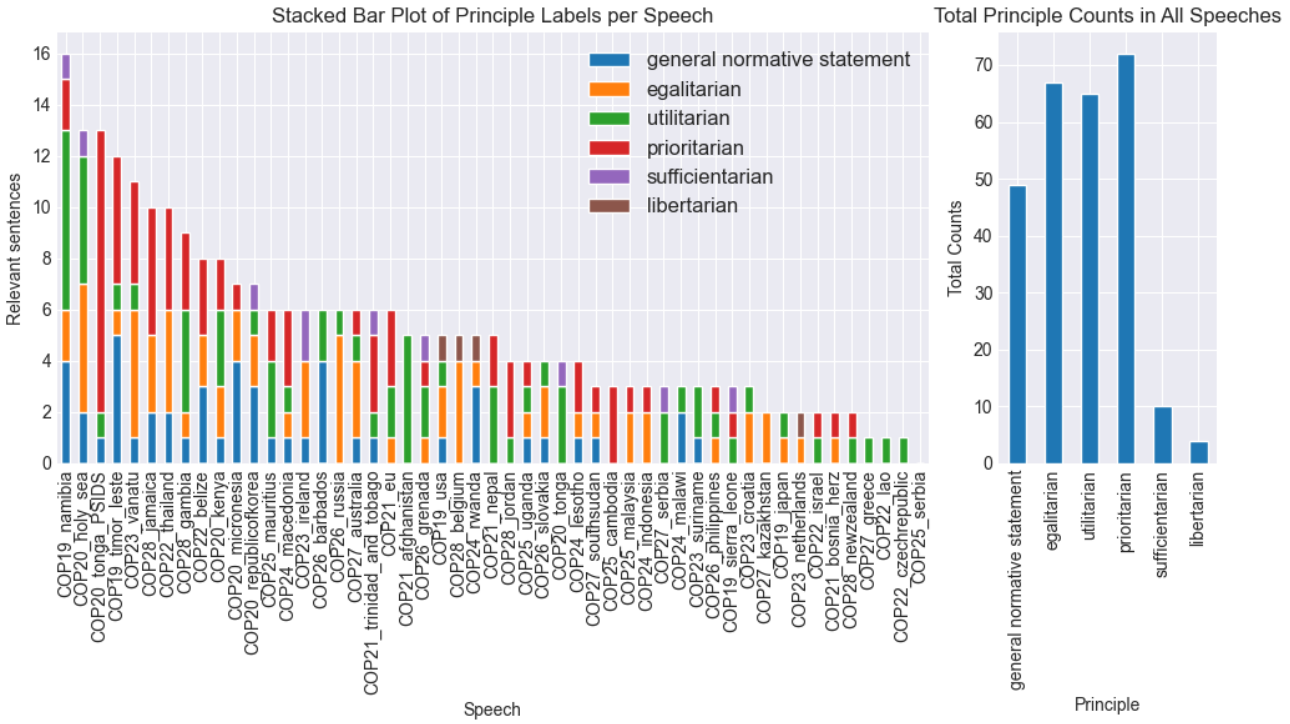

Figure 10 Manual annotation - Principle labels - distribution per speech and over the full dataset

Figure 11 visualizes dominant principle in in all evaluated speeches. The most dominant principle per speech is Utilitarian, dominant in fifteen speeches. This contrasts with the overall most dominant principle; Prioritarian. All but one of these utilitarian dominant speeches are presented by non-Annex I countries. In twelve speeches, no dominant principle is found. The egalitarian principle is dominant in twelve speeches, seven of which are Annex-I. Seven countries dominantly present the prioritarian principle. All prioritarian dominant speeches are presented by a non-Annex I country, except the EU. The sufficientarian and libertarian principle are not the dominant principle in any speech. Micronesia, The Republic of Korea, Rwanda, and Malawi most dominantly present general normative statements that cannot be tied to a single principle.

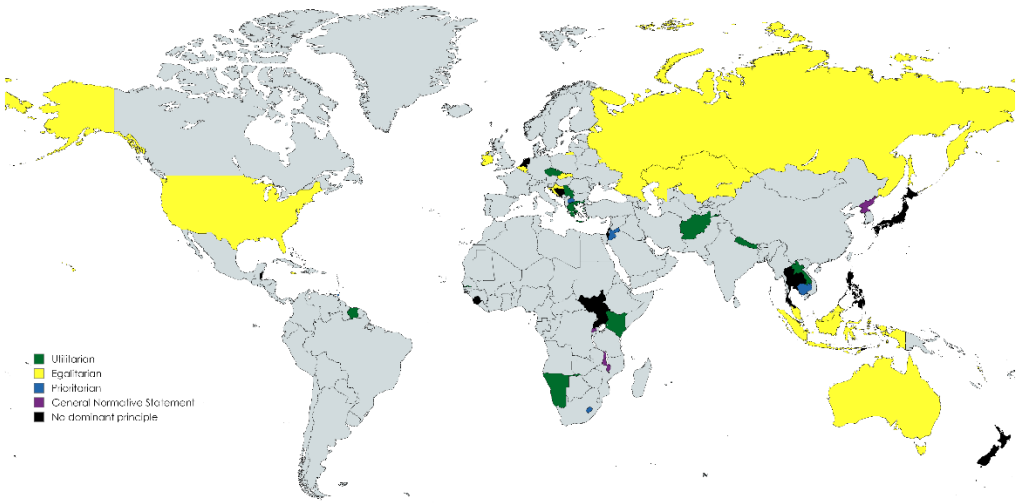

Figure 11 Manual annotation - Coloured map of most dominant principles in annotated speeches per country. The utilitarian principle is most dominant globally. Note that this dominance is only determined for a single speech for a specific COP.

4.2.3 | Categories

Topic labels reveal that limited consistency is found in the subjects that are discussed in speeches. More than 20% of all relevant sentences are labelled as ‘other’. Other topics that were found originally were referring to implementation, commitments, gender equality or ocean observation. The main topics highlighted are ones expected in the context of COP; discussions on new UNFCCC policies and discussions on previous agreements. The topic urgency is highlighted in 15% of the sentences, cooperation in 10% of the sentences. A preliminary observation can be made that non-Annex I countries more often discuss the topic of financial mechanisms.

The unit of the distribution is indicated in 55% of the relevant sentences. The responsibility unit is discussed in the context of urgency and cooperation, indicating that accounting for climate change is a shared responsibility. It is mentioned a context of differing responsibilities, referencing the differences between Annex I and non-Annex I countries and the responsibility of the developed countries to help others. After responsibility, the unit of financial resources is indicated most often, followed by technological, and general support units, indicating the large role of redistribution of resources in the context of climate change.

In 60% of sentences, the preferred shape could not be derived. The priority to the worst off is most often indated, followed by proportional to commitments. A strong correlation is found between the principle prioritarian principle, the units financial resources and support, the topic of financial mechanisms, and the shape “priority to worst off”. Correlation is also found between the utilitarian principle and shapes proportional to commitment and the topic of urgency. Shapes equity and equality show correlation with both the egalitarian and utilitarian principle. No significant correlations are found in relation to the sufficientarian and libertarian principle.

In Appendix E, full distribution of all categories over the speeches is presented as well as a correlationmatrix presenting all correlations between labels that reach a 0.5 treshold.

4.3 | Main findings

The manual annotation revealed that labelling for distributive justice principle is a complex process where edge cases and annotator interpretations are highly influential. The first interpretational challenges arise when labelling sentences for relevance. The inclusion of a Statement of Intent label gives room to label both edge cases and policy intentions presented. Informed judgement is used to label relevant sentences with principles, but semantics still present room for interpretation. Similar word choices can be found in sentences labelled with the same principle. The topic of the sentence is labelled to give insights into the context in which principles are evoked. The occurrence of ‘other’ labels indicates that the use of a pre-definied set of Topic labels compromises the insights that can be gained. The dimensions unit and shape can be related to the practical implementation of a principle, as presented by a country. These are not always specified and cannot be identified in almost half of the relevant sentences. Five codebooks, each covering the labels for a single category, have been created.

The ground truth dataset includes 267 relevant sentences which are annotated for principles and three categories that provide contextual information. Exploratory data analysis revealed that the characteristics of speeches differ per presenting country. Large class imbalances and correlations between the elements have been identified, with the sufficientarian and libertarian principle only identified a limited number of times. The prioritarian principle is most dominant when aggregating all speeches. Aggregating per country, the utilitarian principle is most dominant. Some correlations are found between category labels and the prioritarian, egalitarian and utilitarian principle. Dimensions related to financial distributions are mainly found in relation with the prioritarian principle. Most relevant sentences do not present a specific shape, but when they do, most often prioritizing the worst off is indicated. It is concluded that category labelling can be beneficial to understand the context of distributive justice evocations in texts. However, by only evaluating the categories, principles cannot be determined, indicating that more, non labelled, elements are influential in the annotation process.

5 | Part B: LLM annotation

The ground truth dataset and codebooks created during the manual annotation are used to evaluate the performance of GPT-4o in distributive justice evaluation tasks. This chapter presents the results of both the primary and secondary experiments.

5.1 | Results – Primary experiments

The created ground truth dataset was split into a test and train set containing 20% and 80% of the data. A split is made based on full speeches, selecting one random speech per COP. The context-based evaluation limits the use of random sampling methods like k-fold cross-validation. The train dataset consists of 1212 sentences, the test dataset consists of 331 sentences. In Appendix F class balance of both test and train datasets are presented. This split was originally made to evaluate potential overfitting of the model, for example due to the use of specific prompt examples. Each model-prompt-combination was first applied on the train dataset and model performance was evaluated. This was compared to the test set performance.

Each prompt is evaluated with five instances of the model. To all instances, post-call filtering for the same fingerprint is applied. This results in up to five annotations for each sentence. As omitted batches of sentences are not similar per seed, only sentences that are annotated five times are selected. Predictions are evaluated per seed and consistency in predictions over five seeds is determined. Ultimately, an aggregated dataset of predictions is created, selecting the majority label in case of inconsistencies. For this dataset, the performance of the model-prompt combination is evaluated for performance and incorrect classifications.

5.1.1 | B1 Pre-processing

The pre-processing task labels all sentences for relevance. Runtime for the train dataset annotation with five seeds is 20 minutes. Table 8 presents the number of annotated sentences, omitted batches and inconsistently annotated sentences for each of the prompts.

Accuracy for all prompts is over 55% on the train set. It increases with the inclusion of examples for both datasets, with a maximum of 6% increase in case of context-based instructions. The addition of these instructions does not improve performance. Performance on the test set is generally worse, with accuracy about five percent lower than on the training set, but a similar increase in performance can be found when examples are added to the prompt.

Table 8 B1 Preprocessing – Annotated consistently annotated and evaluated sentences of the train set. Accuracy variance between seeds is <2%

Prompt	Missed batches	Nr. of sentences 5x annotated	% of total	Nr. of inconsistently annotated sentences	% inconsistent of 5x annotated	Accuracy
B1.0	12	972	80%	109	11%	0.60
B1.0.1	14	940	78%	97	10%	0.56
B1.1	22	860	71%	72	8%	0.63
B1.1.1	13	952	79%	118	12%	0.62

As the pre-processing task is created with the aim to use it to identify relevant sentences, performance metrics are valued differently per label. Table 9 and 10 present the aggregated metrics for all prompts. Ultimately it is seen that the addition of examples and context instructions improve performance slightly. The results indicate that recall of Relevant sentences is high for all prompts, but a trade-off is made with regards to precision. This leads to the observation that there is a large number of false positives. The support column indicates the number of sentences of the ground truth dataset that were used to calculate performance metrics.

Identification of Not relevant labels reveal a precision of up to 96%. Inclusion of examples or context leads to a small decrease in performance. Recall of not relevant sentences is +- 50%, revealing that only half of true not-relevant sentences are identified as such. Adding examples improves the recall on the train dataset, where performance on the test set remains more consistent.

The more ambiguous label of Statement of Intent presents a less distinct Precision and Recall trade-off, with both metrics being around 50% for all prompts. This indicates that less than half of the sentences is correctly labelled as statement of intent, and almost half of the true statements of intent are missed. This performance is consistent over the four prompts and the addition of examples does not improve performance.

Table 9 B1 Preprocessing – Train set
The best performing prompt, based on F1 score and support sentences, for each label is made bold.

Label: Relevant				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.34	0.81	0.48	171
B1.0.1	0.32	0.81	0.46	170
B1.1	**0.37**	**0.82**	**0.51**	**146**
B1.1.1	0.35	0.84	0.49	162
Label: Not relevant				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.96	0.55	0.70	642
B1.0.1	0.94	0.50	0.65	601
B1.1	**0.94**	**0.61**	**0.74**	**582**
B1.1.1	0.94	0.58	0.72	613
Label: Statement of intent				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.46	0.56	0.51	159
B1.0.1	0.46	0.52	0.49	169
B1.1	0.43	0.52	0.47	132
B1.1.1	**0.51**	**0.54**	**0.53**	**177**

Table 10 B1 Preprocessing – Test set
The best performing prompt, based on F1 score and support sentences, for each label is made bold.

Label: Relevant				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.27	0.71	0.39	35
B1.0.1	0.23	0.68	0.34	41
B1.1	0.27	0.67	0.38	39
B1.1.1	**0.30**	**0.73**	**0.42**	**49**
Label: Not relevant				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.86	0.57	0.69	160
B1.0.1	0.90	0.47	0.62	177
B1.1	0.88	0.55	0.68	190
B1.1.1	**0.88**	**0.58**	**0.70**	**212**
Label: Statement of intent				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B1.0	0.49	0.44	0.46	55
B1.0.1	0.46	0.48	0.47	52
B1.1	0.45	0.56	0.50	61
B1.1.1	**0.51**	**0.51**	**0.51**	**69**

Figure 12 presents the confusionmatrix for the train set of B1.1, presenting insights into the incorrect annotation. This figure shows that relevant sentences are most often mislabelled as statements of intent. This can be expected as Statement of intent is a label for which edge cases are found. Not relevant statements are also mostly mislabelled as statements of intent. True statements of intent are only once labelled as not relevant, all other sentences were incorrectly labelled as relevant. It is important to highlight that 15% (21 sentences) of relevant sentences are labelled as not relevant. This is of larger influence as not relevant sentences are expected to be of no value in the determination of distributive justice principles and are not further evaluated.

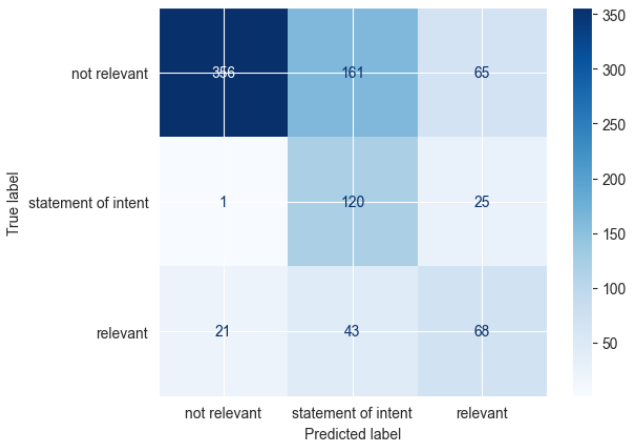

Figure 12 B1 Preprocessing - Confusionmatrix B1.1 Train set

5.3.2 | B2 Principles – Relevant sentences

This task labels the relevant sentences of the train and test set for principles. Only in the evaluation of the first prompt, three batches are omitted. All other tests, both for the test and train dataset, evaluate all sentences. Runtime for train set annotation with five seeds is 3 minutes.

All accuracy is above 50% and limited differences can be found between the accuracy of the train sets. However, it must be taken into account that prompt B2.0 is not evaluated for all sentences, potentially increasing the accuracy. Annotations of the test set are less accurate than the train set, but the addition of examples increases accuracy with 10%. As can be seen in Table 11, less than 15% of sentences is annotated inconsistently, which is slightly higher than the relevance annotation task.

Table 11 B2 Principles - Annotated consistently annotated and evaluated sentences of the train and test set. Accuracy variance between seeds is <2%

Prompt	Dataset	Missed batches	Nr. of sentences 5x annotated	% of total	Nr. of inconsistently annotated sentences	% inconsistent of 5x annotated	Accuracy
B2.0	Train	3	160	73%	23	14%	0.68
	Test	0	49	100%	5	10%	0.51
B2.1	Train	0	218	100%	23	11%	0.67
	Test	0	49	100%	4	8%	0.61

Table 12-14 present the performance metrics for all principles in both the train and test sets. The performance differences between B2.0 and B2.1 are highlighted again, where a more significant difference can be found: an almost 20% increase in macro average recall with the inclusion of examples. The three most dominant principles – Prioritarian, Utilitarian, and Egalitarian – are identified well, reaching F1 scores of 70% in with both prompts. Their dominance can be partly seen as explanation for this performance, where the model simply has more data. Slight improvement with the prioritarian label, is found when examples are added. The test set performance on utilitarian is low compared to the train set performance, potentially due to class imbalances.

Within the annotation process of relevance, class imbalances have a higher influence, with limited availability of libertarian and sufficientarian labels. Within the train set, confusion can be found between prioritarian, sufficientarian, and utilitarian. Libertarian principles are not correctly identified with B2.0 in both test and train data. The only true libertarian label in the train set has been labelled as utilitarian, which is seen in the correlation matrix presented in Figure 13. Prompt B2.1 annotates two sentences as libertarian, of which one is correct. This can be seen in the correlation matrix of Figure 14. The sufficientarian sentence has not been correctly identified in any of the test sets.

Table 12 B2 Principles - B2.0 Performance on each label.- Train set
Labels are sorted on best F1 performance

Label	Precision	Recall	F1	Support
Prioritarian	0.79	0.74	0.76	46
Utilitarian	0.63	0.83	0.72	41
Egalitarian	0.75	0.63	0.68	43
Sufficien-tarian	0.40	0.67	0.50	6
GNS	0.56	0.39	0.46	23
Libertarian	0	0	0	1
Macro avg.	0.52	0.54	0.52	160

Table 13 B2 Principles – B2.0 Performance on each label.- Test set
Labels are sorted on best F1 performance

Label	Precision	Recall	F1	Support
Prioritarian	0.67	1	0.80	6
Egalitarian	0.73	0.69	0.71	16
GNS	0.80	0.25	0.38	16
Utilitarian	0.22	0.67	0.33	6
Sufficien-tarian	0	0	0	2
Libertarian	0	0	0	3
Macro avg.	0.40	0.43	0.37	49

Table 14 B2 Principles - B2.1 Performance on each label.- Train set

Labels are sorted on best F1 performance

Label	Precision	Recall	F1	Support
Prioritarian	0.81	0.82	0.81	66
Utilitarian	0.67	0.73	0.70	59
Libertarian	0.50	1	0.67	1
Egalitarian	0.71	0.57	0.63	51
Sufficien-tarian	0.46	0.75	0.57	8
GNS	0.45	0.42	0.44	33
Macro avg.	0.60	0.71	0.64	218

Table 15 B2 Principles – B2.1 Performance on each label.- Test set

Labels are sorted on best F1 performance

Label	Precision	Recall	F1	Support
Prioritarian	0.67	1	0.80	6
Egalitarian	0.69	0.69	0.69	16
Libertarian	1	0.33	0.50	3
GNS	0.45	0.42	0.44	33
Utilitarian	0.6	0.67	0.47	6
Sufficien-tarian	0	0	0	2
Macro avg.	0.57	0.53	0.51	49

With prompt B2.1, egalitarian statements are most often misclassified as general normative statements. Additional confusion is found between classifications of general normative statements and utilitarian statements. In both prompts, one true egalitarian statement is identified as libertarian. This is surprising due to their conceptual differences. No other sentences are wrongly predicted as libertarian, indicating that this definition has limited overlap with other principles. The matrices reveal that the prioritarian and sufficientarian labels can be confused. In both datasets, two true sufficientarian sentences are labelled as prioritarian, and three true prioritarian sentences are labelled as sufficientarian. The in-practice implementation of these two principles, often focussing on helping the worst off, could be an explanation.

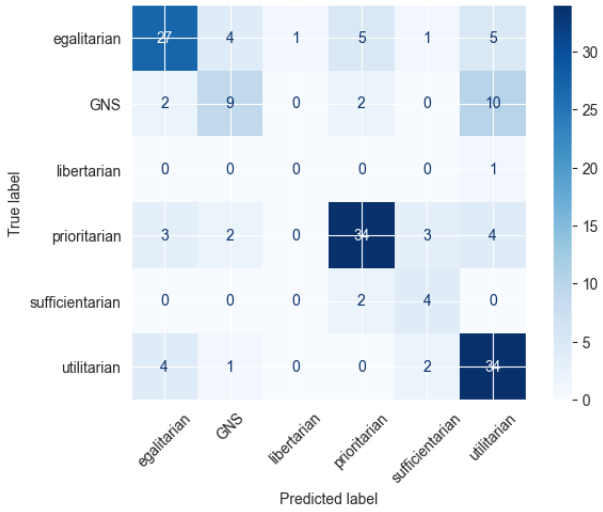

Figure 13 B2 Principles - B2.0 Confusionmatrix Train data

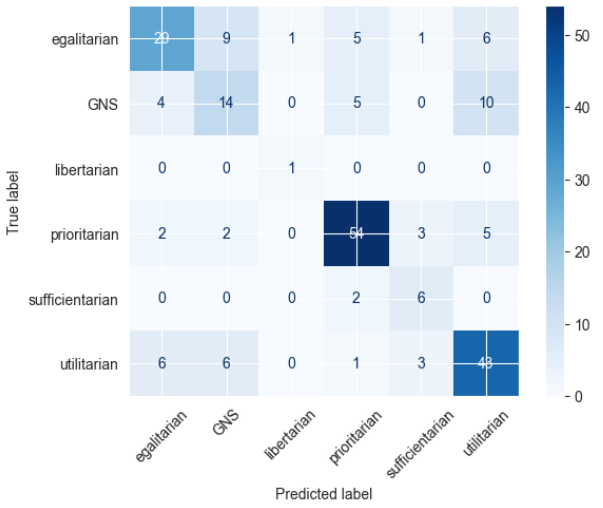

Figure 14 B2 Principles - B2.1 Confusionmatrix Test data

Overall, it is seen that although the accuracy does not differ much between the train and test set, the performance metrics of annotations with prompt B2.1 are slightly higher. The rare categories sufficientarian and libertarian are identified better with addition of examples. The category General Normative Statements presents the most misclassifications, but this can be expected due to the ambiguity of this category.

5.3.3 | B3 Principles –all sentences

Task B3 identifies distributive justice principles the full training dataset of 1212 sentences, without any pre-processing of relevance. Same class imbalances in the relevant sentences apply, and a total of 992 not relevant sentences, consisting of the ground truth statement of intent and not relevant sentences, are included in the Train set. The test set contains 281 not relevant sentences. The not-evaluated label is the biggest class in both datasets. Runtime for train set annotation with five seeds is 20 minutes.

Table 16 presents the missed batches, annotated sentences and accuracy of both the train and test sets for all prompts. The number of missed batches in the B3.1 test evaluation and B3.1.1 train evaluation. This leads to a smaller number of sentences that are annotated five times. The number of annotated sentences is in case of the test set for prompt B3.1 only 57% of all sentences. The percentage of inconsistently annotated sentences decreases in the test set with the addition of examples, but percentage of inconsistently labelled sentences remains stable. Accuracy for all prompt evaluations is over 60%, with a small increase when examples are added to the prompt. However, the large class imbalances make this number deceiving and performance for each label is evaluated.

Table 16 B3 Preprocessing – Annotated consistently annotated and evaluated sentences of the train and test set. Accuracy variance between seeds is <2%

Prompt	Dataset	Missed batches	Nr. of sentences 5x annotated	% of total	Nr. of inconsistently annotated sentences	% inconsistent of 5x annotated	Accuracy
B3.0	Train	4	1132	93%	171	11%	0.61
	Test	11	251	76%	38	15%	0.61
B3.0.1	Train	13	1068	88%	159	15%	0.60
	Test	13	310	93%	43	14%	0.60
B3.1	Train	6	1172	97%	130	11%	0.64
	Test	20	190	57%	33	17%	0.69
B3.1.1	Train	21	832	69%	93	11%	0.66
	Test	11	250	76%	42	17%	0.69

Tables 17 and 18 show the performance for both test and train datasets for the four different prompts. The best performing prompt for each label is made bold. Best performance is based on the F1 score and the largest number of support sentences. Evaluating the macro average performance for both the train and test set indicates that performances does not reach satisfactory F1 scores. Only the Recall comes close to the 0.5 threshold. The identification of not-relevant sentences, labelled as not-evaluated, is done well by the model. Almost 100% of non-relevant labels are applied to true non-relevant sentences and more than 60% of sentences is identified.

Not evaluated labels are dominant in the dataset and performance on all other, non-dominant classes, is less satisfactory. Egalitarian, Prioritarian, and Utilitarian labels are identified with a recall mostly above 60%, but precision reaches a 40% maximum. Sufficientarian labels are identified in the train set, with recall over 50%, but are not found by the model in the test set. The limited occurrence of the libertarian sentences, combined with the choice to omit batches leads to no libertarian sentences in the evaluation the train set with prompt B3.0.1 and B3.1.1. Additionally, this label was not identified in the test set. Performance for this label can thus not be accounted for. In case of prompt B3.1, five sentences are identified as libertarian, with one of these predictions correct and four others having the true label Not Evaluated. Sufficientarian labels are mostly identified correctly, with only one misidentification as prioritarian.

The identification of General Normative statements proves to be complex when non-relevant sentences are included in the dataset. Especially precision is far below the 0.5 threshold. Confusionmatrices, like Figure 15, reveals that sentences that are labelled as Not Evaluated in the ground truth dataset, are often predicted as General Normative statement, Utilitarian, or Prioritarian. This indicates that the model deems more sentences to present normative statements. As most sentences are labelled as General Normative Statements, it can be interpreted that the model does not find a principle that matches. This also works the other way around, with sentences that have principles as true label and are labelled as GNS by the model. Examples are Utilitarian sentences. These sentences are mislabelled 19 times as GNS. Although the definition of Utilitarian is provided, the model seems to interpret these sentences as having some sort of motivation but failing to classify it as utilitarian.

Table 17 B3 Principles – Train set**.** The best performing prompt, based on F1 score and support sentences, for each label is made bold.
**Libertarian ground truth label was omitted in post-call filtering*

Label: Egalitarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.39	0.61	0.48	49
B3.0.1	0.29	0.60	0.39	43
B3.1	**0.41**	**0.60**	**0.48**	**52**
B3.1.1	0.38	0.65	0.48	31
Label: Prioritarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.35	0.66	0.46	68
B3.0.1	0.29	0.68	0.41	59
B3.1	**0.38**	**0.62**	**0.47**	**65**
B3.1.1	0.30	0.49	0.37	49
Label: Utilitarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.26	0.64	0.35	50
B3.0.1	0.25	0.58	0.35	57
B3.1	0.28	0.54	0.36	56
B3.1.1	**0.37**	**0.67**	**0.48**	**39**
Label: Sufficientarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.36	0.71	0.48	7
B3.0.1	0.28	0.62	0.38	8
B3.1	**0.46**	**0.86**	**0.60**	**7**
B3.1.1	0.25	0.60	0.35	5
Label: Libertarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0	0	0	1
B3.0.1	*0*	*0*	*0*	*0**
B3.1	0.20	1	0.33	1
B3.1.1	*0*	*0*	*0*	*0**
Label: General Normative statement				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.10	0.64	0.17	33
B3.0.1	0.09	0.41	0.14	32
B3.1	0.11	0.76	0.19	34
B3.1.1	**0.13**	**0.70**	**0.22**	**27**
Label: Not evaluated				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.99	0.60	0.75	930
B3.0.1	0.98	0.61	0.75	869
B3.1	0.99	0.65	0.78	957
B3.1.1	**0.98**	**0.68**	**0.80**	**681**
Macro avg.				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.35	0.55	0.38	1132
B3.0.1	0.31	0.50	0.35	1068
B3.1	**0.40**	**0.72**	**0.46**	**1172**
B3.1.1	0.34	0.54	0.39	832

Table 18 B3 Principles – Test set**.** The best performing prompt, based on F1 score and support sentences, for each label is made bold.
Within the test set, no sufficientarian and libertarian labels have been identified.

Label: Egalitarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	**0.43**	**0.90**	**0.58**	**10**
B3.0.1	0.36	0.71	0.48	14
B3.1	0.39	0.88	0.54	8
B3.1.1	0.40	0.77	0.53	13
Label: Prioritarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.42	0.83	0.56	6
B3.0.1	0.28	0.83	0.42	6
B3.1	**0.42**	**1**	**0.59**	**5**
B3.1.1	0.33	0.83	0.48	6
Label: Utilitarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.13	0.50	0.21	6
B3.0.1	0.08	0.5	0.14	6
B3.1	**0.20**	**0.33**	**0.25**	**6**
B3.1.1	0.07	0.20	0.11	5
Label: Sufficientarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0	0	0	1
B3.0.1	0	0	0	2
B3.1	0	0	0	2
B3.1.1	0	0	0	1
Label: Libertarian				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0	0	0	1
B3.0.1	0	0	0	3
B3.1	0	0	0	1
B3.1.1	0	0	0	3
Label: General normative statement				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.11	0.54	0.19	13
B3.0.1	0.09	0.36	0.14	14
B3.1	**0.18**	**0.64**	**0.29**	**11**
B3.1.1	0.16	0.50	0.24	12
Label: Not evaluated				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	1	0.61	0.76	214
B3.0.1	0.98	0.62	0.76	265
B3.1	1	0.71	0.83	157
B3.1.1	**0.97**	**0.72**	**0.83**	**210**
Macro avg.				
Prompt	**Precision**	**Recall**	**F1**	**Support**
B3.0	0.30	0.48	0.33	251
B3.0.1	0.25	0.43	0.28	310
B3.1	**0.31**	**0.51**	**0.36**	**190**
B3.1.1	0.28	0.43	0.31	250

Following the performance metrics, it is seen that performance improves with the addition of examples. Including context instructions can minor further improvement on the F1 score, but most of the time a different trade-off between precision and recall is found.

Figure 15 B2 Principles - Confusionmatrix B3.1

5.3.4 | B4 Characteristics

Task B4 instructs the model to label sentences for Principle and the categories Topic, Unit, and Shape. It is performed both on the full dataset, including not relevant sentences and Statements of intent, and on a subset of only the relevant sentences. Although all codebooks include instructions on the output space, these are not always followed by the GPT. For this multiclass multilabel classification task, labels of one category (e.g. relevance) are used in another (e.g. Topic). For evaluation purposes sentences with incorrect predicted class labels are removed from the outcome dataset.

Only 644 sentences (53%) of the full train dataset have complete annotations for prompt B4. 379 of these sentences present some inconsistency in one of the categories over five seeds. Annotation of the full train set is subject to the same class imbalances as classification task B3, and class imbalances are increased by the limited number of complete annotations. Contrary to B3, performance on the 'not evaluated' label does not reach annotation thresholds. This label is found in the ground truth dataset of all of all categories but is not included in any predictions other than the Principle category. Overall accuracy of these predictions is very low, as shown in Table 19, for both the test and train dataset.

Table 19 - B4 Categories – Accuracy for Principle, Topic, Unit, and Shape.

Category	Accuracy Train – Full	Accuracy Test – Full	Accuracy Train - Relevant
Principle	0.22	0.23	0.55
Topic	0.15	0.15	0.53
Unit	0.21	0.18	0.74
Shape	0.17	0.16	0.59

Evaluation of a subset of relevant sentences, thus excluding the 'not evaluated' label in all categories, leads to a significant change in performance. 194 of 331 sentences are fully annotated, 86 with inconsistency in at least one of the categories. In the test set consisting of 49 relevant sentences, only 18 are fully annotated, with 6 presenting inconsistencies. This dataset is so limited that evaluation

of performance is not possible. This is why only the performance on the relevant sentences of the train dataset is evaluated in more detail. Performance metrics for each of the four categories and their respective labels are presented in Table 20-23.

Table 20 B4 Categories - Label Principle – Performance Relevant Train dataset

	Precision	Recall	F1	Support
egalitarian	0.53	0.65	0.58	48
GNS	0.35	0.19	0.25	31
libertarian	0	0	0	1
prioritarian	0.83	0.54	0.65	56
sufficientarian	0.31	0.71	0.43	7
utilitarian	0.53	0.69	0.60	51
accuracy			0.55	194
macro avg	0.43	0.46	0.42	194

Performance on the train dataset generally presents good performance for the prioritarian, utilitarian and egalitarian labels. Specifically, prioritarian sentences are identified with high precision. General normative statements are not identified correctly, with low recall and precision. The one libertarian sentence in the dataset is not identified by the model as such.

Table 31 B4 Categories - Label Topic – Performance Relevant Train dataset

	Precision	Recall	F1	Support
UNFCCC agreements and principles	0.67	0.36	0.47	28
adaptation	0.17	1	0.29	1
adaptation and mitigation	0.63	0.56	0.59	9
cooperation	0.39	0.85	0.53	20
financial mechanisms	0.62	1	0.76	21
mitigation	0.31	0.67	0.42	6
new UNFCCC policy	0.55	0.39	0.45	31
other	0.74	0.33	0.45	43
urgency	0.58	0.54	0.56	35
accuracy			0.53	194
macro avg	0.51	0.63	0.50	194

Most topics are identified with performance metrics below the minimum threshold, but the 100% recall and 62% precision of the financial mechanism contributes to a higher overall performance. The trade-off between precision and recall can again be identified. Precision of the adaptation, cooperation, and mitigation topics is below the 50% threshold. These do present high recall, indicating that more sentences are labelled with these topics than that are found in the ground truth dataset. On the other hand, UNFCCC related topics and sentences that cannot be appointed a specific topic are identified with better precision and lower recall. Performance of UNFCCC related topics can be explained by their more solid definitions, with words directly pointing to UNFCCC principles. The topics urgency and the combined adaptation and mitigation present a less clear trade-off.

Table 22 B4 Categories - Label Unit – Performance Relevant Train dataset

	Precision	Recall	F1	Support
financial and technological resources	0.46	0.86	0.60	7
financial resources	0.97	0.97	0.97	30
not indicated	0.85	0.74	0.79	99
other	0	0	0	3
responsibility	0.59	0.71	0.64	41
support	0.50	0.50	0.5	12
technological resources	0.50	1	0.67	2
accuracy			0.74	194
macro avg	0.55	0.68	0.60	194

The Unit category reaches, as only category, a positive sufficiency of 0.7 with an accuracy of 0.75. Almost all instances of financial resources are correctly identified. This is probably due to the clear notion of finance related words in these sentences. Additionally, the model performs well on identifying what sentences present a unit, as the Not Indicated label reaches an F1 score of 0.79. The recall of the combined financial and technological resources is high, indicating that these true labels are found. This recall is set against a precision slightly below 50%. The unit Responsibility shows a similar performance. Only three sentences were originally labelled as presenting a different type of unit. These instances were not captured by the model.

Table 23 B4 Categories - Label Shape – Performance Relevant Train dataset
**The 'not evaluated' category should have been excluded from the dataframe. Two sentences have been mislabelled and this has not been corrected.*

	Precision	Recall	F1	Support
equality	0.18	0.67	0.28	9
equity	0.43	0.50	0.46	12
needs based	0.18	0.75	0.29	4
not evaluated*	0	0	0	2
not indicated	0.90	0.60	0.72	120
priority to worst off	0.50	0.50	0.50	26
proportional to commitment	0.61	0.69	0.65	16
proportional to contribution	0.80	0.80	0.80	5
accuracy			0.59	194
macro avg	0.45	0.56	0.46	194

Finally, the category Shape. The model correctly classifies most instances in which a shape is not presented. In addition, the shapes proportional to commitment and proportional to contribution are well classified. Especially proportional to contribution is identified well. This is probably due to the use of the word 'contribution' in these sentences. Priority to worst off and Equity present a less clear trade off between precision and recall, leading to F1 scores around 50%. Precision of Equality and Needs based labels is very low, indicating that more sentences are classified with this label than there actually are.

Task B4 is the most elaborate of the four and this is seen in its performance. Instructions are not followed correctly, and results have to be filtered. Annotation of the full dataset led to extremely low accuracy, indicating that the model was not able to annotate correctly. Accounting for the Not Evaluated label was not correctly interpreted, leading to accuracy below 20%. Performance on the train dataset of relevant sentences provided sufficient outcomes to come to evaluation. A clear trade-off could be found between precision and recall, but as the main focus is to come to correct classifications, the combined metric F1 can be seen as most important.

5.4 | Results – Secondary experiments

Four secondary experiments have been performed to evaluated influence of model settings on GPT performance and annotations. These experiments have been performed for task B1, determining relevance. The different prompts used can be found in appendix G. Performance metrics have been calculated the same way as the primary experiments.

An increase in temperature to 0.6 for five seeds leads to an increase in inconsistently annotated sentences. For evaluation of codebook B1.0, 30% of fully annotated sentences present inconsistencies. This is almost three times as many as with a temperature of 0. Performance increased slightly, with accuracy of 0.64 compared to 0.60.

As described, several precautions have been taken to ensure that annotations are reproducible and consistent. Experiment S2, performed with three iterations of the five different seeds and a temperature of 0, revealed that inconsistencies do still occur. In 10% of sentences that were annotated for three iterations, inconsistencies in predictions were found. Performance, when selecting the

majority label for final evaluation, is similar to the original experiment. Accuracy is 60%, with performance metrics for all labels within 2% range of the original experiment. Combining both a temperature of 0.6 and three iterations in S3 resulted in annotations with slightly more inconsistent sentences compared to S2. Again, no significant change in performance was found.

Finally, a binary annotation was performed with an altered version of GPT_annotate. Most important finding of this annotation process is that annotations created by the model include more than one annotation per row. Consistency in labelling is also limited. Codebook alterations had no effect on this phenomenon.

5.4 | Main findings

Four annotation tasks have been performed with GPT-4o. Measures implemented to adapt to the non-deterministic nature of the model lead to batches being omitted during the annotation task. No consistency can be found in the number of missed batches. Only sentences that are annotated five times are used to calculate performance metrics. On average, 12% of these sentences are not labelled the same by all five model instances. This inconsistency increases substantially when the temperature of the model is increased to 0.6. Over all experiments, including exploratory research, less than 75 dollars was spend.

It is found that performance metrics improve when one-shot prompts are used. The addition of context instructions to one-shot prompts incidentally leads to minimal performance improvement. This improvement mainly found either one of the datasets. Only annotating for Statement of Intent, in the pre-processing task, seems to present a 0.06 and 0.01 improvement on the train and test set respectively.

The pre-processing task indicates a precision-recall trade-off for both relevant and not relevant labels. Recall is high (>0.8) for the relevant sentences, but precision is limited (<0.4). The identification of not relevant sentences done with high precision (>0.9) and lower, but satisfactory, recall (>0.5). The Statement of Intent label is identified correctly in about half the time, with precision and recall around 0.5.

Identifying the distributive justice principles in sentences that are deemed relevant, task B2, is performed with 60% accuracy. Large class imbalances influence model performance. The more dominant labels – Utilitarian, Egalitarian, and Prioritarian – are identified all performance metrics around 0.7. The rarer Sufficientarian and Libertarian cases are not always identified. Combining both the pre-processing and principle identification tasks indicates that the model reaches high precision in determining what sentences are Not Evaluated, indicating that they are either not relevant or Statements of Intent. The model does label more sentences as relevant than there originally are, most of these are incorrectly labelled as General Normative Statements. The precision of the assigned principles proved to be limited and under 0.4.

The final multiclass-multilabel annotation task proved to be complex as labels were used for non-corresponding classes. Accounting for the Not Evaluated label was not correctly interpreted, leading to accuracy of annotations on the full dataset below 20%. Only the annotation of relevant sentences of the train set resulted in sufficient annotations for analysis. Principles were identified with 55% accuracy, only slightly lower than the single-class annotation task. Topic identification was performed with 53% accuracy. Best performance was found in identifying financial mechanisms. The unit was performed with 74% accuracy, with especially high performance on the unit of financial resources and sentences that do not clearly indicate a unit. Identification of shapes is done with 59% accuracy. Again, sentences in which this category is not indicated is labelled well. The shapes proportional to contribution is identified best, with precision, recall, and F1 score of 0.8.

6 | Discussion

This study has presented insights into GPT-4o performance on multiple distributive justice annotation tasks. To reach these insights a ground truth dataset was created of 51 manually annotated HLS speeches. The results contribute to the two aims of this research: descriptive knowledge generation and prescriptive insights for both modelers and policymakers. This discussion section evaluates limitations of both the manual and LLM annotation. Directions for future work are presented as well as a more fundamental discussion on the concept of identifying normative motivations in text.

6.1 | Manual annotation

The manual annotation process was funded on the theoretical conceptualisation of distributive justice. Here five principles where identified. Four of these principles present clear distributional preferences, prescribing a distribution to reach equality or to benefit all. The libertarian principle can be seen as an odd one out in this approach. This principle is funded in the idea of freedom, mostly preferring minimal redistribution and focussing on market-based solutions. It can be argued that this makes it not a true principle of distributive justice. However, the principal stance against (forced) redistribution makes it relevant to evaluate as it takes a stance against the other principles. The current selection of principles could be elaborated upon. Speeches have presented elements of *limitarian* ideas, for example the Holy Sea prescribing the need to change our consumption-based society. Other distributive justice theories than the currently defined ones could be present in the speeches, however it is expected that the egalitarian, utilitarian and prioritarian principles will remain the most common.

Key empirical reflections in relation to the manual annotation include the notion of textual ambiguities and the unavoidable interpretational choices and considered judgements that must be made. This was first found in the need to pre-process sentences for relevance. Not all sentences point to a distributive justice principle, but it is complex to determine when it does. This research determines relevance based on normativity, but this adds another level of subjective interpretation. A future research direction would be to label groups of sentences or paragraphs for principles. This would enable to take more context into account and potentially reducing ambiguities. However, new complexities would arise to determine to what extent sentences are connected.

The complete sentences and their context are of influence on the principle that is identified. This identification is not always straightforward, as motivations are not noted. The diplomatic nature of the text leaves even more room for interpretation. For this reason, the ‘general normative statement’ label was introduced. However, this label does not account for sentences where multiple motivations can be found. The manual annotation process could be improved by re-evaluating these sentences to see if there is no principle identifiable or that they can be interpreted in multiple ways. Following this interpretation path, it would be beneficial for the quality of the ground truth dataset to have a second annotator annotate the relevant sentences. This annotator would have access to the complete theoretical foundation as outlined in this research and is tasked to label based on this information and the provided codebooks. It is expected that a limited number of sentences will be annotated differently. These edge cases can then be highlighted and evaluated. This does not mean that one of the interpretations is correct, but that these sentences are expected to also be edge cases in LLM annotations.

The labels used in the Topic category can be studies to understand the connection between certain topics and principles. An example is that sentences that discuss the topic of urgency correlates with the utilitarian principle. Less significant correlations were found than initially expected, however, more with evaluation of more speeches this could change. With this evaluation it is advised not to use pre-determined topics, as this takes away from the diversity of elements that are discussed and might present a limited view on the relationship between principles and the context in which they are evoked.

The Unit and Shape categories are notion of the practical implementation of distributive justice. It is found that these dimensions are not always indicated in sentences but if they are, they present additional insights on how principles are used. Some empirical observations were made in the differnce between notion of these dimensions by Annex-1 and non-Annex 1 countries. This indicates that altough same principles are indicated, views on practical implementation can differ. The found correlation between financial dimensions and the principle of prioritarianism was expected. The correlation between the utilitarian principle and the topic of urgency and shape proportional to commitment were more surprising. More detailled evaluation of these connections is a recommended area of future work. Along the lines of this work it would be of value to study the similarities between the distributive justice principles mentioned in relation to new policies and in relation to policies that countries already implemented or intent to impliment in their home countries. A subsection of these policies is already mentioned in the same speeches and labelled as Statement of Intent. Additionally the NDC's and NC's could also be evaluated. It is expected that dominant principles per country, as the scope diverts from global to national, changes.

Patterns
It is found that all principles are indicated in speeches, the prioritarian principle is most found. However, when evaluating per speech, utilitarian and egalitarian principles are the most often the dominant principle. The empirical explanation for this pattern would be that the prioritarian principle is evoked in a larger number of speeches, but mostly for a single sentence, often referring to financial mechanisms. It must be noted that this observation has only been made on a subset of speeches and can thus not be seen as fully representative. However, it is a first indication of the broad moral plurality that is found in climate policymaking. This again highlights that with respect to this level of policymaking, moral plurality must be considered. Only accounting for the dominant principles would disregard moral values of a group of parties. Finally, some patterns could be identified between the dominant principles found in annex-1 and non-annex 1 countries. Future research directions could take these insights and, if the dataset is increased, study the connection between used principles and vulnerability to climate consequences.

6.2 | LLM annotation

From the outset, a severe limitation is the non-deterministic nature of this model. The biggest challenge is the non-controllability of the system_fingerprint, which can change during the annotation process and over the course of days. As post-call filtering is performed, annotations are discarded, and it is suspected that performance varies over different fingerprints used. The non-deterministic nature was confirmed by performing multiple iterations of an annotation task for one model instance. Mostly deterministic outcomes were expected for these three iterations, however, inconsistencies remained. The post-call filtering, combined with the large class imbalances, leads to an additional implication, where rare labels like libertarian sentences, being discarded. This complicates performance evaluation where the number of labels available in the dataset influence performance metrics.

As GPT models do not learn from data used in API calls, the split into test and train data was not of the significant value. Essentially, both a larger and smaller dataset of completely new speeches were evaluated. This split is not necessary in future research that does not contain a learning element in the LLM used. However, a true LLM classifier could benefit from learning the diversity of sentences that can be found in these speeches.

Performance
General performance over all codebook types is moderately consistent but can differ significantly per label. The importance of the pre-processing task is highlighted by the performance difference between task B2 and B3. Performance without pre-processing leads to many misclassified sentences, disregarding ground truth instances principles. It is important to note that instructions to include context do not result in significant improvement in performance. For this reason, it is recommended to perform the pre-processing task before labelling for principles. The context-based instructions could

be of limited added value because the batches of sentences were pre-split and did not cover a full speech. Future work could evaluate the effect of adding a full speech and instructing the GPT to split the text and evaluate the sentences. It is expected that context instructions will have more influence.

Some principles are better identified than others. Utilitarian, Egalitarian, and Prioritarian are identified all performance metrics around 0.7. It is theorized that, as these principles are more found in the texts, they are better identified. Libertarian and Sufficientarian sentences were not always identified, but this finding also indicates that false positives were also not found. This indicates that the GPT did not find more cases than during manual annotation. A potential direction for future work following from these findings would be to instruct the GPT to label on elimination basis, presenting what principles the sentence does not indicate. This could also benefit the identification of edge cases.

Most misclassifications are found in more ambiguous categories like statement of intent and general normative statements. In these cases, it is challenging that the GPT does not provide additional reasoning as to why labels are appointed. This is again due to the black-box nature of the model. This could be accounted for by enabling the labelling for multiple principles or asking the GPT to provide a confidence interval. Sentences where the GPT is not confident in the appointed labels could then be manually re-evaluated, preferably by multiple annotators. This to enable evaluation of inter-personal agreement on the labels of the ground truth dataset.

The multiclass/multilabel task proved to be complex and resulted in unstructured and incorrect outcomes. This was not evaluated in full detail and causes could be both instructions based or related to the data processing of GPT output. The task could be performed in parallel with the principle identification. Alternatively, the context categories could be identified separately and used as additional information for the GPT when labelling for principles. In the determination of topics, it is advised to let the model determine topics inductively, as pre-determined labels can only capture the complexity and diversity of texts in a limited way. This would give more room to grasp the range of topics and context in which the principles are evoked in these global negotiations. Findings of this inductive topic-labelling should be compared to topics found with topic-modelling over all speeches, this to evaluate if the context in which distributive justice principles are mentioned are similar to general topics noted in HLS speeches.

Generalizability

Finally, the generalizability of the presented results needs to be discussed. These results are generated with GPT-4o, which is trained on a broad, but unknown dataset. Its performance cannot be directly expected from other generative LLMs, as their performance influenced by its knowledge and understanding of the task. However, definitions are already presented in the prompts and the theoretical foundations described in this research could also be used to perform additional training.

The ground truth dataset can be used to train other LLMs, like ClimateBERT, on a downstream sentence-based classification task. This model would not be instructed by a prompt but uses supervised learning to finding patterns in the provided dataset and potentially with other macro parameters like annex status or vulnerability indexes. Preferably, this model is open-source and can be used to evaluate new texts. This creates a range of scientific possibilities ranging from evaluation of the full corpus of HLS speeches to other types of climate negotiation related texts. Additionally, the option to review non-English speeches should be evaluated. This because availability of English speeches is limited for some parts of the world and potentially the textual context in which principles are presented differs. Extending annotation to non-English texts is essential for broader applicability and reducing language biases.

Other directions of future work would be to train a model on the ground truth dataset and evaluate its performance on other types of justice related texts. This could for example be documents related to the energy transition, where the distributive justice principles are relevant in relation to energy justice. Preferably, these new texts can be evaluated without the need to create a manually annotated subset

first. For this reason, the explainability of annotations should first be improved to enable human annotators to focus on the evaluation of edge cases.

6.3 | Fundamental discussion

This research aims to classify implicitly mentioned motivations and makes a distinction between facts and normative statements. The question “what principle is presented?” can have a seemingly factual answer, i.e. “egalitarian”. However, this answer holds the perspective of the annotator on “what it means to be egalitarian”. This research has aimed to capture standardize these perspectives by providing definitions tailored to the classification task, linking them to global climate policymaking.

The normative nature of the labelling task makes it impossible to create a dataset that is a 100% the ground truth. Subjectivity and interpretational differences can still lead to disagreements on labels, especially as the evaluated texts are written and present diplomatic language and open to interpretation. The categories shape, topic, and unit may sometimes be more objective, but these labels are not sufficient to make a full judgement on the used principle. Human judgement is holistic and like any normative judgement, cannot be simply decomposed in factual predictions and rule applications (Balagopalan, 2023).

Models like GPT-4o present biases we do not know or do not fully understand (Ouyang et al., 2023). Using them for classification tasks that are fundamentally normative can give a false sense of objectivity, especially when it follows a factual evaluation structure (Balagopalan, 2023). For this reason, it can be argued that the identification of distributive justice principles is a human problem that should not necessarily have a technical solution. On the other hand, it is impossible to ask every participant in complex negotiations or other instances of normative uncertainties for their exact thoughts, motivations and values that influence their options. The application of LLM based principle evaluation should thus be done with caution and with the awareness of the limitations of the models used.

7 | Conclusion

This research has explored the use of LLMs for identifying distributive justice preferences. A subset of COP HLS speeches was manually annotated, creating a ground truth dataset and annotation codebooks. GPT-4o was presented with multiple classification tasks, both multi-class and multi-label, and performance was evaluated against the created ground truth dataset. Technical and fundamental limitations of this research have been identified. To conclude, the sub-questions are presented and synthesized to answer the main research question:

To what extent do Large Language Models accurately identify distributive justice preferences in climate negotiation texts?

1. What are the key theories and concepts of distributive justice relevant in the context of climate negotiations?

Practical use of distributive justice covers three dimensions: scope, unit and shape. Principles of distributive justice are used as moral guidelines and prescribe different preferences with regards to the dimensions. Five key principles have been identified and defined in the context of climate policy. Egalitarianism advocates for reducing inequalities and ensuring fairness in the distribution of resources, opportunities, and responsibilities. Utilitarianism prioritizes maximizing overall welfare or benefits, disregarding individual, or subgroup differences. Prioritarianism emphasizes increasing the welfare of the worst off, aiming to uplift those in the most disadvantaged positions. Sufficientarianism focuses on ensuring that everyone reaches a minimum threshold of well-being, while libertarianism emphasizes individual freedom and opposes forced redistribution.

Principles can be found in various elements of global climate negotiations, from the mostly egalitarian foundations of the UNFCCC and the principle of CBDR, to the sufficientarian principles of human rights and needs. Different distributive justice principles are most explicitly noted in the context of financial mechanisms and emission rights, where distributions are subject to large stakes and societal consequences.

2. Which distributive justice principles and preferences are expressed by parties in HLS speeches?

All key principles of distributive justice are found in HLS speeches. The theoretical framework formulated by answering the first sub question was used to evaluate and annotate a subset of 51 HLS speeches from 10 different COPs. It was found that principles are mostly implicitly implied. To filter for relevant sentences, a pre-processing evaluation task is created. In this task a distinction has been made between relevant (normative) sentences and not relevant (factual) sentences. Because principles are seen motivational, it is assumed they cannot be implied in factual statements. The middle-ground label ‘statement of intent’ covers sentences of which normativity can be disputed, for example when policy intentions are outlined. Even with pre-processing for normativity, some sentences do not present indication of a single distributive justice principle. These sentences are labelled as ‘general normative statements’.

In addition to labelling for relevance, sentences are annotated for three contextual elements. First is the topic of the sentence, second the implied distributive unit and finally the implied preferred distributional shape. Textual ambiguities create room for different interpretations and robustness of the dataset could be improved by employing a second annotator. Ultimately, five codebooks are created, one for each category, including label definition funded in theory and empirical findings.

The created ground truth dataset includes 267 relevant sentences which are annotated for principles and three categories that provide contextual information. Exploratory data analysis revealed that the characteristics of speeches differ per presenting country. Large class imbalances and correlations between the elements have been identified, with the sufficientarian and libertarian principle only identified a limited number of times. The prioritarian principle identified the most often in all evaluated

speeches. The utilitarian principle is however the most dominant when aggregating principles per speech.

3. How does LLM identification of distributive justice preference compare to human annotated data?

GPT-4o performance on identification of distributive justice preferences shows, compared to human-annotated data, mixed performance with notable strengths and limitations. Four annotation tasks were conducted evaluating performance on pre-processing tasks, identifying distributive justice principles in both the full dataset and pre-processed dataset, and identifying not only the principles, but also the additional categories Topic, Unit, and Shape.

Only sentences annotated with each of the five model instances were used to calculate performance metrics. Inconsistencies were found in these annotations, with on average 12% of sentences presenting inconsistent labels. In the pre-processing task, the model showed a precision-recall trade-off: high recall (>0.8) but limited precision (<0.4) for Relevant sentences, and high precision (>0.9) and satisfactory recall (>0.5) for Not relevant sentences. This indicates that more sentences are deemed relevant by the model than by a human annotator, but when a sentence is labelled as Not relevant, it is most likely to be correct.

Identifying distributive justice principles in relevant sentences was performed with 60% average accuracy. In the identification of Utilitarian and Prioritarian sentences did not present a precision-recall trade-off and were identified with 0.8 and 0.7 recall and precision. Combining both the pre-processing and principle identification tasks indicated that the model performs well identifying sentences that were not to be labelled for principles. However, the model does label more sentences as relevant than the ground truth dataset, most of these are incorrectly labelled as General Normative Statements. The precision of the assigned principles proved to be limited and under 0.4, indicating that principles were not correctly identified. The combined annotation task, essentially performing the same annotations as the manual annotation, did not present promising results. Category labels were mixed up and Not relevant sentences were not accurately identified. Annotating relevant sentences for the four categories reached more promising results. Principles were identified with 55% accuracy. Topic identification was 53% accurate, best for financial mechanisms (74%), while shapes and units had varying accuracies.

Overall, annotation of relevant sentences reached 60% accuracy, exceeding 70% for certain principles with examples included. Without pre-processing, the model showed high precision for not-relevant sentences but less accuracy for principles, often mislabelling not-relevant sentences. Despite good precision in some areas, the LLM struggled with overall accuracy and consistency compared to human annotations. However, the misclassifications and most misclassifications are found in more ambiguous labels which are also complex for human annotators to evaluate.

4. What are the strengths and limitations of LLM application for the distributive justice identification task?

A clear advantage of employing GPT-4o for annotation purposes is that it significantly reduces time investments. Evaluation of 41 speeches with five model instances took less than 20 minutes or about 30 seconds per speech. This is a fraction of manual annotation time which took 20 minutes per speech. The model's performance underlines its potential, both for pre-processing and identification purposes. Additionally, costs are limited as all experiments were performed for a total of less than 75 dollars. However, the non-deterministic nature of the model combined with the large number of edge cases lead to inconsistencies in appointed labels, which must be re-evaluated by a human. Additionally, the black box nature of the model potentially leads to biases currently unknown. Using these models for classification tasks that are fundamentally normative can give a false sense of objectivity.

To finally answer the main research question, the generalizability of this research needs to be addressed. The generalizability of the results generated with GPT-4o, trained on an unspecified

dataset, is limited to this model and may not directly apply to other generative LLMs. However, the theoretical foundations and provided definitions in the research can facilitate additional training. The extent to which all LLMs can accurately identify distributive justice principles and preferences is up to debate. However, this research has shown its potential in at least supporting this annotation task by accurately performing pre-processing tasks.

7.1 | Scientific contributions

The aim of this research was to generate descriptive knowledge about the use of distributive justice principles and present prescriptive insights for both modelers and policymakers. This knowledge was gained in the context of one of the most complex policymaking arenas: the UNFCCC COP. To account for the normative uncertainties found in climate policymaking, a deep understanding of the various perspectives and moral justifications is necessary. This research addresses this understanding by using a theoretical foundation in a bottom-up approach, identifying the distributive justice principles used by parties in these negotiations.

This research has indicated that a bottom-up analysis is possible in this context. The analysis of sentences has proved that distributive justice principles can be identified in HLS speeches. This adds to the existing research on climate justice and its role at COP. In addition to these principles, characteristics of practical implementation – Shape and Unit – can be identified in a structured and streamlined way. This broadens the understanding of the ways parties prefer principles implemented in future policies. The bottom-up identification process is time-consuming when performed manually but is divided by 40 when annotating with GPT-4o. Although human evaluation will always be necessary, this insight can be of value in future evaluation of both climate related text and other context in which distributions can cause normative uncertainties, like the energy transition, even if only the model's pre-processing abilities are used.

The ground truth dataset is one of the few labelled datasets that touches on normative evaluations and is, to my knowledge, the only dataset that performs these evaluations on HLS speeches. Subjectivity and interpretational differences can still lead to disagreements on labels, but it can be seen as a starting point for research in this area. Although the ground truth dataset is only a subset of all HLS speeches presented during COP19 to COP28, some initial patterns are found that prove to be insightful in both policymaking and modelling. The insight that all principles are found in HLS speeches confirms the moral plurality of the policymaking and indicates that only accounting for a single principle disregards this diversity. It is also found that the globally most used perspective – the prioritarian principle – is not the most dominant when evaluating on country level. This indicates that simply selecting a majority principle would also be a compromising choice. The identified dimension Shape presents additional insights on the varying distributional shapes that are preferred. A surprising distributional shape is "proportional to commitments". This principle is different from traditional SWFs and adds to the debate on the use and approach to these functions as this sheds a different light on preferred distributions.

Overall, it is confirmed that an arbitrary selection of distributive justice principles disregards the diversity of moral guidelines used in the climate debate and that nuanced bottom-up understanding can create new perspectives on current modelling practices.

7.2 | Societal relevance

This research highlights the complexity of distributive justice and its influence in the climate domain. Its insights and bottom-up approach contribute to the understanding of normative uncertainties. This can lead to new policies that are perceived as fair, or at least fairer, by a broader range of parties. In the broad sense, this research contributes to the understanding of moral foundations of national positions in the climate debate, presenting room for more inclusive and constructive dialogues as insights are gained as to why certain policies are proposed.

This study demonstrates the potential and limitations of using NLP and specifically GPT-4o when evaluating complex normative statements. Although these models can save a lot of time, the main take-away for society is here that these models do not present objective annotations. Results generated with these models should always be subject to a form of human evaluation, especially when there is room for interpretation.

Ultimately, this research highlights the fact that some principles are dominant in the climate debate. This does not indicate that these principles are the most ‘right’. The presence of minority perspectives, like the sufficientarian principle, should not be disregarded in the climate debate. Majority ideas can easily overshadow the dialogue and it is increasingly important compare and evaluate personal values and priorities to the values presented by decisionmakers.

7.3 | Future directions

The potential research directions following this research are broad and range from training different types of LLMs to evaluating other types of texts for distributive justice principles. Nevertheless, further research should start with improving some fundamental elements.

A critical next step would be to focus on the explainability of the LLM annotations. Research should be directed towards developing methods and tools that provide clear and interpretable explanations for the model's classifications. This will contribute to the transparency of the method, allowing researchers to understand and validate the AI's annotations. In addition, this could point to edge cases in the text which then can be evaluated by multiple annotators. If necessary, the annotation in the ground truth dataset could then be changed, improving its validity and reducing biases.

After improving the explainability of model annotations and validity of the ground truth dataset, the full corpus of HLS speeches can be evaluated. This will enable the full evaluation moral justifications used over 10 years of COP, enabling the study of temporal patterns and other potential macro parameters like vulnerability indexes or GDP. Ultimately, the bottom-up approach to understanding normative uncertainties can always be expanded with new data, new principles, and even new dimensions of justice.

8 | Reflection

The initial goal of this research was to produce work that would not only advance my own knowledge but also contribute to addressing a global challenge. This ambition was underpinned by grand research ideas: not only identifying distributive justice principles using GPT and analysing which principles were predominantly used by countries over the course of 10 COPs, but also evaluating these principles against IAM modelling practices. The aim was to see how real-life use of distributive justice principles was reflected in modelling; a gap that had been identified by Jafino et al. (2021). However, the challenge of performing such broad research within a limited timeframe soon became apparent.

The first major challenge was the availability of speeches and incomplete metadata in the UNFCCC database. Collecting them in one go was impossible, necessitating the scraping of web pages. Converting PDFs to text presented additional difficulties. Some documents were poorly scanned and not directly convertible. Additionally, inconsistencies in layout—such as varying text formats and margins—further complicated the process, often resulting in entire paragraphs not being converted accurately. These issues were resolved by manually checking all text documents and removing unnecessary text, such as logos. I sincerely hope that this corpus will be used by others creating it was extremely time consuming. The texts themselves are fascinating, offering rich insights into the diverse perspectives of countries on climate change and the types of policies they prefer. However, there remains a gap between rhetoric and action, which I recommend future research should explore.

A significant underestimation on my part was time management. Understanding complex subjects takes time, and it is easy to get lost in the philosophical rabbit holes, continually reading papers on new perspectives. The biggest challenge was determining the endpoint of my research. This was particularly true during manual annotation, where it was tempting to continually refine and iterate. Nonetheless, I have gained a profound understanding of philosophy, justice, and the limitations of GPT—knowledge that is invaluable in our increasingly AI-driven society.

I have greatly enjoyed discussing the implications of this research with others and have learned that the struggle with diverging opinions makes this research relatable. I hope this research will be valuable to policymakers, helping them understand that intrinsic motivations can differ yet lead to similar policy preferences. Whether driven by prioritarian or utilitarian motivations, the outcome can often be the same: a commitment to helping people.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anadkat, S. (2023) How to make your completions output consistent with the new seed parameter. Retrieved from: https://cookbook.openai.com/examples/reproducible_outputs_with_the_seed_parameter

Arneson, Richard, "Egalitarianism", The Stanford Encyclopedia of Philosophy (Summer 2013 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2013/entries/egalitarianism/>.

Adler, M., Anthoff, D., Bosetti, V. *et al.* Priority for the worse-off and the social cost of carbon. *Nature Clim Change* 7, 443–449 (2017). https://doi.org/10.1038/nclimate3298

Bagozzi, B. E. (2015). The multifaceted nature of global climate change negotiations. *Review of International Organizations*, *10*(4), 439–464. https://doi.org/10.1007/s11558-014-9211-7

Balagopalan, A., Madras, D., Yang, D. H., Hadfield-Menell, D., Hadfield, G. K., & Ghassemi, M. (2023). Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances*, *9*(19), eabq0701.

Biesbroek, R., Wright, S. J., Eguren, S. K., Bonotto, A., & Athanasiadis, I. N. (2022). Policy attention to climate change impacts, adaptation and vulnerability: a global assessment of National Communications (1994–2019). Climate Policy, 22(1), 97-111.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. https://arxiv.org/pdf/2005.14165

Cairney, P., Timonina, I., & Stephan, H. R. (2023). How can policy and policymaking foster climate justice? A qualitative systematic review. Open Research Europe, 3, 51. https://doi.org/10.12688/openreseurope.15719.2

Caney, Simon, "Climate Justice", The Stanford Encyclopedia of Philosophy (Winter 2021 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2021/entries/justice-climate/>.

Chancel, L., & Piketty, T. (2015). Carbon and inequality: From Kyoto to Paris. https://doi.org/10.13140/ RG.2.1.3536.0082

Davidson, M. D. (2021). How fairness principles in the climate debate relate to theories of distributive justice. *Sustainability (Switzerland)*, *13*(13). https://doi.org/10.3390/su13137302

Ding, B., Qin, C., Liu, L., Chia, Y. K., Joty, S., Li, B., & Bing, L. (2022). Is gpt-3 a good data annotator?. *arXiv preprint arXiv:2212.10450*.

Driver, Julia, "The History of Utilitarianism", The Stanford Encyclopedia of Philosophy (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/win2022/entries/utilitarianism-history/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

European Environment Agency [EEA]. (2024, March 25). Climate change mitigation: reducing emissions. Retrieved May 2, 2024, from https://www.eea.europa.eu/en/topics/in-depth/climate-change-mitigation-reducing-emissions#:~:text=Mitigating%20climate%20change%20means%20reducing,important%20part%20of%20the%20solution.

Füssel, H.-M. (2010). How inequitable is the global distribution of responsibility, capability, and vulnera- bility to climate change: A comprehensive indicator-based assessment. *Global Environmental Change*, *20*(4), 597–611. https://doi.org/10.1016/j.gloenvcha.2010.07.009

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30),

Green Climate Fund (n.d.). About GCF. Retrieved from https://www.greenclimate.fund/about

Huseby, R. (2020). Sufficiency and the threshold question. The Journal of Ethics, 24(2), 207-223.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research*, *15*(9), 1277-1288.

Islam, S. N., & Winkel, J. (2017). *Climate Change and Social Inequality* * (tech. rep.). https://www.un.org/esa/desa/papers/2017/wp152_2017.pdf

Jafino, B. A., Kwakkel, J. H., & Taebi, B. (2021). Enabling assessment of distributive justice through models for climate change planning: A review of recent advances and a research agenda. https://doi.org/10.1002/wcc.721

Kartha, S., Athanasiou, T., Caney, S. *et al.* Cascading biases against poorer countries. *Nat. Clim. Chang.* 8, 348–349 (2018). https://doi.org/10.1038/s41558-018-0152-7

Kolstad C., K. Urama, J. Broome, A. Bruvoll, M. Carino Olvera, D. Fullerton, C. Gollier, W. M. Hanemann, R.Hassan, F. Jotzo, M. R. Khan, L. Meyer, and L. Mundaca*, 2014: Social, Economic and Ethical Concepts and Methods. In: Climate Change2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*

Konow, J. (2003). Which is the fairest one of all? A positive analysis of justice theories. *Journal of economic literature*, *41*(4), 1188-1239.

Lamb, W. F., Mattioli, G., Levi, S., Timmons Roberts, J., Capstick, S., Creutzig, F., Minx, J. C., Müller- Hansen, F., Culhane, T., & Steinberger, J. K. (2020). Discourses of climate delay. *Global Sustain- ability*, *3*. https://doi.org/10.1017/sus.2020.13

Lamont, Julian and Christi Favor, "Distributive Justice", The Stanford Encyclopedia of Philosophy (Winter 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2017/entries/justice-distributive/>.

Lykkeskov, A., & Gjerris, M. (2017). The Moral Justification Behind a Climate Tax on Beef in Denmark. *Food Ethics*, *1*(2), 181–191. https://doi.org/10.1007/s41055-017-0017-1

MacQueen, K. M., McLellan, E., Kay, K., & Milstein, B. (1998). Codebook development for team-based qualitative analysis. *Cam Journal*, *10*(2), 31-36.

Meyer, Lukas & Roser, Dominic (2006). Distributive Justice and Climate Change. The Allocation of Emission Rights. *Analyse & Kritik* 28 (2):223-249.

Newell, P., Srivastava, S., Naess, L. O., Torres Contreras, G. A., & Price, R. (2021). Toward transformative climate justice: An emerging research agenda. *WIREs Climate Change*, *12*(6). https://doi.org/10.1002/wcc.733

Nordhaus, W. (2018). Evolution of modeling of the economics of global warming: changes in the DICE model, 1992–2017. *Climatic Change*, *148*(4), 623–640. https://doi.org/10.1007/s10584-018-2218-y

Okereke, C. (2010). Climate justice and the international regime. Wiley interdisciplinary reviews: climate change, 1(3), 462-474. https://doi.org/10.1002/wcc.52

OpenAI (2024), Hello GPT-4o, retrieved from https://openai.com/index/hello-gpt-4o/

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023). LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation. *arXiv preprint arXiv:2308.02828*.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative ai requires validation. arXiv preprint arXiv:2306.00176.

Pottier, A., Méjean, A., & Godard, O. (2017). A Survey of Global Climate Justice: From Negotiation Stances to Moral Stakes and Back. *International Review of Environmental and Resource Economics*, *11*(1), 1–53. https://doi.org/10.1561/101.00000090

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086. https://doi.org/10.1038/s41598-024-56706-x

Robeyns, Ingrid. 2017. “Having Too Much.” In Wealth: NOMOS LVIII, edited by Jack Knight and Melissa Schwartzberg, 1–44. New York: New York University Press.

Robeyns, I. (2019). What, if anything, is wrong with extreme wealth?. *Journal of Human Development and Capabilities*, *20*(3), 251-266.

Rubiano Rivadeneira, N., & Carton, W. (2022). (In)justice in modelled climate futures: A review of integrated assessment modelling critiques through a justice lens. *Energy Research & Social Science*, *92*, 102781. https://doi.org/10.1016/j.erss.2022.102781

Sahoo, B. K., & Murari, K. K. (2023). Development of an Integrated Assessment Model in the Climate Policy Framework and Its Challenges. https://doi.org/10.1007/978-981-99-1388-6{_}23

Schlosberg, D., & Collins, L. B. (2014). From environmental to climate justice: climate change and the discourse of environmental justice. *WIREs Climate Change*, *5*(3), 359–374. https://doi.org/10. 1002/wcc.275

Sietsma, A. J., Ford, J. D., & Minx, J. C. (2024). The next generation of machine learning for tracking adaptation texts. *Nature Climate Change*, *14*(1), 31–39. https://doi.org/10.1038/s41558-023-01890-3

Sinnott-Armstrong, Walter, "Consequentialism", The Stanford Encyclopedia of Philosophy (Winter 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/win2023/entries/consequentialism/>

Stede, M., & Patz, R. (2021). *The Climate Change Debate and Natural Language Processing* (tech. rep.).

Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise?. *arXiv preprint arXiv:2306.13906*.

Sikdar, S.K. On aggregating multiple indicators into a single metric for sustainability. *Clean Techn Environ Policy* 11, 157–161 (2009). https://doi.org/10.1007/s10098-009-0225-4

Taebi, B., Kwakkel, J. H., & Kermisch, C. (2020). Governing climate risks in the face of normative uncertainties. *WIREs Climate Change*, *11*(5). https://doi.org/10.1002/wcc.666

Törnberg, P. (2023a). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Törnberg, P. (2023b). How to use LLMs for Text Analysis. http://arxiv.org/abs/2307.13106

United Nations Secretariat (UN/DESA) [UN]. (2014). Country classification. In World Economic Situation and Prospects (p. 144). https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf

UNFCCC. (n.d.) *Party Groupings.* Retrieved from: https://unfccc.int/process-and-meetings/parties-non-party-stakeholders/parties/party-groupings

UNFCCC. (1992). *United Nations Framework Convention on Climate Change* . retrieved from: *https://unfccc.int/resource/docs/convkp/conveng.pdf*

UNFCCC. (2020). How to COP. Retrieved from: https://unfccc.int/sites/default/files/resource/How-to-COP_2020.pdf

UNFCCC. (2022) Reference Manual for the Enhanced Transparency Framework under the Paris Agreement. Retrieved from: https://unfccc.int/sites/default/files/resource/v2_ETFreferencemanual.pdf

United Nations (2023) List of least developed countries. Retrieved from: https://www.un.org/development

Van de Poel, I., & Royakkers, L. (2023). Ethics, technology, and engineering: An introduction. John Wiley & Sons.

van der Vossen, Bas and Billy Christmas, "Libertarianism", The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/libertarianism/>

Van Uffelen, N., Taebi, B., & Pesch, U. (2024). Revisiting the energy justice framework: Doing justice to normative uncertainties. *Renewable and Sustainable Energy Reviews*, *189*, 113974.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). *CLIMATEBERT: A Pretrained Language Model for Climate-Related Text* (tech. rep.). www.github.com/climatebert/language-model

Weyant, J. (2017). Some contributions of integrated assessment models of global climate change. *Review of Environmental Economics and Policy*, *11*(1), 115–137. https://doi.org/10.1093/reep/rew018

Wright, S. J., Sietsma, A., Korswagen, S., Athanasiadis, I. N., & Biesbroek, R. (2023). How do countries frame climate change? A global comparison of adaptation and mitigation in UNFCCC National Communications. *Regional Environmental Change*, *23*(4), 129. https://doi.org/10.1007/s10113- 023-02113-3

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023, March). Supporting qualitative analysis with large language models: Combining codebook with GPT-3 for deductive coding. In *Companion proceedings of the 28th international conference on intelligent user interfaces* (pp. 75-78).

Zimm, C., Mintz-Woo, K., Brutschin, E., Hanger-Kopp, S., Hoffmann, R., Kikstra, J. S., Kuhn, M., Min, J., Muttarak, R., Pachauri, S., Patange, O., Riahi, K., & Schinko, T. (2024). Justice considerations in climate research. *Nature Climate Change*, *14*(1), 22–30. https://doi.org/10.1038/s41558-023- 01869-0

Appendix A | Data preparation

This study has created a corpus of 1083 PDF documents of speeches presented during the HLS of COP19 (2013) to COP28 (2023). The documents were scraped from their corresponding COP HLS page of the UNFCCC database on 25/03/2024. The PDF documents were parsed to extract the core texts. A manual review was performed to make sure there were no conversion problems and any non-core elements were removed.

The original languages of the texts are Arabic, French, Spanish, and Russian. Of the 198 parties, 47 have no written English speeches available (23%). Although translation and analysis by LLMs are possible, this research is scoped to only evaluate English written texts. This decision was made to stay as close to the original as possible, with translation potentially disregarding secondary meanings of words, influencing interpretations.

Figure 1 visualizes the countries with limited to no availability of English speeches. This research is scoped to only evaluate the 743 English written texts. The restriction excludes almost all of Middle America, South America, and Western Africa. These areas are predominantly Spanish, French, and Arab speaking, choosing to present their speeches in their mother tongue. It is important to note that speeches are often made in affiliation with or on behalf of negotiation groups. The developing countries work together in *the Group of 77***,** the chair of this group often speak for all 135 members (UNFCCC, n.d.). This means that the views of the countries without available speeches are not completely disregarded.

All scraped and converted documents, both in PDF and TXT files, are publicly available on GitHub.

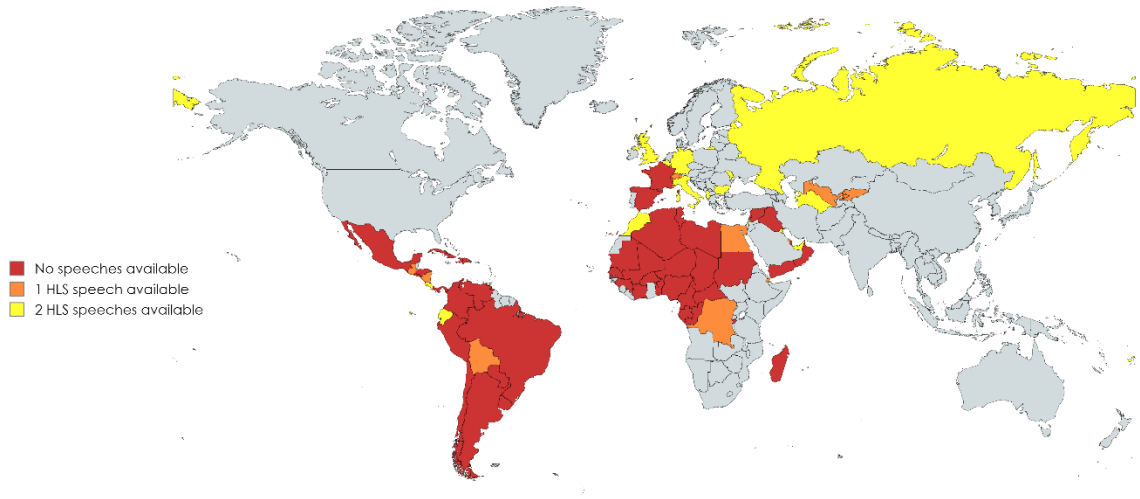

Figure 16 Appendix A - Visualization of English HLS speech availability COP19-COP28

Appendix B | Manual Annotation - Relevance

In this appendix, additional examples on the determination of relevance are presented. The examples presented in Table 2 provide insights in the complexity of this labelling task. The unstructured nature of the sentences, combined with varying context and interpretation bias could lead to other annotators classifying the sentences differently. R3 and R6 are two statements containing expectations for a new agreement, but they are labelled differently due to the additional context they present. Statements R9 could also be interpreted as normative, presenting a normative position on the importance of taking action in light of the moral obligation to protect human lives. Similarly, R8 can be interpreted as normative, if it is evaluated with an emphasis on leaving no one behind. This evaluation focuses on the intent of the implementation of the recovery funds.

Table 24: Appendix B Examples Relevance - Sentences derived from manually annotated HLS speeches.

	(2) Relevant statement	
R1	COP21AFGHANISTAN “I am sure that you all agree with me that we must collectively and responsibly act for the welfare of our common Planet - Terra – a tiny beautiful little blue dot on the Celestial Map of our Solar System which supports Life as we know it.”	This is deemed a normative statement, prescribing collective action and responsibility, motivated by the moral value of protecting the earth and life on it.
R2	COP23VANUATU “The global response to climate change must put fairness and equity at the heart of its work, and to keep the needs and aspirations of the world’s most vulnerable countries in its line of sight."	This statement prescribes the elements that should be considered in new climate policies. The values of fairness, equity, and the protection of the most vulnerable are presented.
R3	COP20REPKOREA “Last but not least, the agreement should also lay solid foundation for transparency and accountability with a clear set of rules.”	This is deemed a normative statement, prescribing the values of transparency and accountability in new policies.
R4	COP19NAMIBIA *“The adverse impacts of climate change know no national boundaries.”* “Global warming is a catastrophic problem that needs a global solution.”	The first sentence is deemed not relevant, prescribing a descriptive statement of the situation. The second sentence is deemed relevant, presenting normative judgement implied with “catastrophic” and prescribing a global solution. This solution is motivated by the previous sentence, indicating that not acting has consequences for all.
	(1) Statement of intent	
R5	COP20MICRONESIA “Mr. President, as we come to the end of the International Year of the SIDS, we hope to leave Lima on a high note.”	This statement presents expectations on the outcome of the negotiations. No normative motivation is presented.
R6	COP21EU “We must translate the momentum we have seen on the road to Paris into an ambitious agreement.”	This statement prescribes the need to take on action, but no normative motivation as to why an ambitious agreement must be reached. Additionally, it is unclear what is indicated with “ambitious agreement”
R7	COP28JORDAN *"This global initiative prioritises climate-related support and investments for refugee -hosting nations."* *"We are grateful for the 58 countries that have supported this initiative so far."* “Much more needs to be done”	The need to take on action is prescribed, indicating an intention to formulate new policies. The previous two sentences (which are classified as not relevant), describe the topic of support in relation to climate migration, but normative motivation is indicated.

R8	COP26SLOVAKIA “43% of our post -pandemic recovery funds will support sustainable green transition in transport, industry or buildings, while not leaving the most vulnerable behind.”	This is a statement of intent denoting the implications of new national policies. The sentence does contain a normative element highlighting the value not to leave the most vulnerable behind, but this is an intent of the prescribed policy.
	(0) Not relevant statement	
R9	COP27SOUTHSUDAN “Therefore, lifesaving interventions are needed urgently else the prospect of climate induced displacement and climate refugees will become a new normal.”	This statement describes the potential consequences of not taking action.
R11	COP26SLOVAKIA "If we fail, the only thing to discuss at future COPs will be the irreversible devastation of our planet and its habitats."	This statement describes the consequences of failing to come to new policies.
R12	COP20REPKOREA " This is crucial to generate confidence on durability and credibility of the new system, not just among parties, but for the “outside world” as well."	This is a descriptive statement that highlights the importance of durability and credibility but does not present any norms or value-based motivation.

Codebook for category Relevance:

Codebook 1: RELEVANCE

Code: Not relevant | *Description:* A sentence that presents a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions

Code: Statement of intent | *Description:* A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.

Code: Relevant | *Description:* A sentence that presents a normative statement with a value judgment or prescription based on norms and values, including motivations derived from the textual context.

Appendix C | Manual Annotation - Principles

Appendix C presents the codebook created for the annotation off five distributive justice principles and general normative statements.

Codebook 2: PRINCIPLE

Code: general normative statement | *Description:* A sentence that presents no indication of foundational distributive justice principle that is used as a motivation for the normative statement.

Code: egalitarian | *Description:* A sentence presenting an egalitarian motivation. Motivated by the goal to reduce inequalities and have equality of opportunity. This includes statements the highlight the inclusion of all and references to mutual contributions by all. The principle of common but differentiated responsibilities is also seen as egalitarian.

Code: utilitarian | *Description*: A sentence presenting a utilitarian motivation. Motivated by the goal of maximizing the benefit of all or contributing to 'the greater good'. The benefit of all can be in present and future generations. Found in sentences prescribing the need to take action, motivated by the need to improve the lives of all.

Code: prioritarian | *Description*: A sentence presenting a prioritarian motivation. Motivated by the moral obligation to help the worst off. Highlighted in sentences focussing on supporting vulnerable developing countries. It does not focus existing differences.

Code: sufficientarian | *Description:* A sentence presenting a sufficientarian motivation. Motivated by the idea that everyone should be able to reach a common level of well-being. Found in sentences referring to a global standard of well-being that everyone should reach as well as statements implying that no one should be left behind this threshold.

Code: libertarian | *Description*: A sentence presenting a libertarian motivation. Motivated by a need for freedom, indicating minimal government involvement and focus on free markets.

Appendix D | Manual Annotation - Categories

Appendix D1 presents the codebooks used in manual annotation of all categories: Topic, Unit and Principle. In Appendix D2 nine sentences with full annotation are presented.

D1 | Category codebooks

Codebook 3: TOPIC

Code: new UNFCCC policy | *Description:* A sentence that discusses requirements for new UNFCCC policy

Code: UNFCCC agreements and principles | *Description:* A sentence that highlight articles, agreements or principles of the UNFCCC

Code: urgency | *Description:* A sentence that discusses the need to take action to address the climate crisis and implement new policies.

Code: cooperation | *Description:* A sentence that focuses on the need for cooperation.

Code: financial mechanisms | *Description:* A sentence that highlights financial redistributions based on a financial mechanism, for example the Green Climate Fund.

Code: adaptation | *Description*: A sentence that focuses on adaptation, adaptation measures, or adaptation policies.

Code: mitigation | *Description:* A sentence that focuses on mitigation, mitigation measures or mitigation policies.

Code: adaptation and mitigation | *Description:* A sentences that highlight both elements of adaptation and mitigation.

Code: other | *Description:* A sentence that discusses a topic that is not covered by any other TOPIC code.

Codebook 4: UNIT

Code: not indicated | *Description:* A sentence that does not highlight a specific unit of distribution.

Code: responsibility | *Description:* A sentence that highlights the distribution of the responsibility to take on action in the context of climate change and climate measures.

Code: financial resources | Description: A sentence that addresses the distribution of financial resources or money.

Code: technological resources | *Description:* A sentence that addresses the distribution of technological resources.

Code: financial and technological resources | *Description:* A sentence that highlights a distribution of both technological and financial resources.

Code: support | *Description*: A sentence that indicates a distribution of support, without specifying what type of support.

Code: other | *Description:* A sentence that discusses a distribution with a unit that is not covered by any other UNIT code.

Codebook 5: SHAPE

Code: not indicated | *Description*: A sentence that does not highlight a specific distributional shape.

Code: equality | *Description*: A sentence that highlights a distributional shape that is equal for all. The size of the allocation does not have to be specified. Sentences highlighting the need for “everyone to take action”, indicating that equality is favoured in the distribution of the obligation to take measures.

Code: equity | *Description*: A sentence that highlights a distributional shape that accounts for differences between the recipients, e.g. accounting for the financial situation of recipients.

Code: priority to worst off | *Description:* A sentence that prescribes a distribution only focussing on distribution to the ones that are deemed to be the worst off, e.g. the distribution of financial resources to countries that are most vulnerable to climate impacts.

Code: needs based | *Description:* A sentence prescribing a distribution that is based on the needs of recipients or what is deemed needed.

Code: proportional to contribution | *Description*: A sentence that indicates a distribution that is based on historical contributions. e.g. in relation to historical responsibility for past emissions, where historically higher polluters have to reduce more emissions.

Code: proportional to commitment | *Description*: A sentence indicating that the distribution should follow the shape of commitments made in global agreements, pledges, and treaties. e.g. found in statements calling upon developed countries to act upon their pledges to commit to the green climate fund.

D2 | Fully annotated sentences

[F1] COP19 Namibia	**The Green Climate Fund, which we regarded as an ideal beginning to address the finance aspect for climate change remains dry.**
Principle: prioritarian **Topic: financial mechanisms** **Unit: financial resources** **Shape: not indicated**	Presents normative judgement on the lack of resources in the Green Climate Fund. This fund is developed to help developing countries in their implementation of climate measures, making it prioritarian in nature. The unit of the distribution is financial resources. The preferred shape, relating to the distribution of financial resources in a financial mechanism, is not specifically indicated. *Not labelled as* responsibility *and proportional to commitment as there is no explicit call to act upon the responsibility of developed countries to adhere to their commitments to the climate fund.*
[F2] COP20 Holy Sea	The longer we wait, the more it will cost; more victims will suffer from our inaction and the greatest weight will fall on the most vulnerable, the poorest peoples and future generations: what is at issue here is respect for their fundamental human rights.
Principle: sufficientarian **Topic: urgency** **Unit: not indicated** **Shape: not indicated**	Presents a normative statement prescribing the need to act. This prescription is motivated by the respect for fundamental human rights. These rights are seen as a global baseline level that everyone should reach, making the motivation sufficientarian. There is no specific distribution prescribed, only the need to take on action.
[F3] COP20 Micronesia	The major polluters, especially those who are most responsible for climate change, have the moral obligation to do much more, to take the lead and to raise their ambition now.
Principle: general normative statement **Topic: UNFCCC agreements and principles** **Unit: responsibility** **Shape: proportional to contribution**	Presents normative judgement on responsibility of major polluters to take on action. This directly tied to article 4 of the convention, common but differentiated responsibilities. It highlights a distribution of responsibility proportional to past contributions to climate change. The motivational principle cannot be derived, leading to the label: general normative statement. *Although CBDR can be seen as an egalitarian motivated principle, this statement does not call for contributions by all, focussing on the major polluters. The statement could also be motivated utilitarian, arguing that that the moral obligation to help would be in the benefit of all. This ambiguity has led to the label: general normative statement.*
[F4] COP23 Croatia	We must activate all our efforts to preserve the environment for future generations, and related to it I would like to thank to all those who, despite the many obstacles and difficulties, continue in these efforts.
Principle: utilitarian **Topic: urgency** **Unit: not indicated** **Shape: not indicated**	Normative statement prescribing the need to take on actions, motivated by the need to protect the environment for future generations. This indicates a utilitarian motivation. No unit or shape of a distribution is indicated, other than a general need to take on action. *This statement expresses both a prescriptive normative statement and expresses gratitude. Due to the normative section, the statement is labelled as relevant.*

[F5] COP23 Netherlands	We should make the most of these opportunities by developing and implementing innovative solutions, enabling our businesses to contribute to, and benefit from, the global transition towards a carbon -neutral economy.
Principle: libertarian **Topic: other** **Unit: not indicated** **Shape: not indicated**	Normative statement with focus on achieving a carbon-neutral economy by supporting businesses. The explicit focus on economics and business is also found in the libertarian principle valuing an economical and market-based approach to new policies. No unit or shape of a distribution are indicated.
[F6] COP23 Vanuatu	All nations must raise their ambition and implement significant sector wide emission reductions as a matter of urgency.
Principle: egalitarian **Topic: mitigation** **Unit: responsibility** **Shape: equality**	Prescriptive normative statement on urgent need to take on action. Relates to a distribution of responsibility to take on this action, this distribution account for all, making it equality based. The notion that everyone should take action is deemed motivated by the egalitarian principle. *The topic of this sentence can also be interpreted as urgency, however the specific notion of urgency to reduce emissions makes this a better fit.*
[F7] COP19 Namibia	As a country committed to address the adverse effects of climate change, Namibia is ready to increase its mitigation efforts provided that sufficient financial and technical support is provided.
Principle: sufficientarian **Topic: mitigation** **Unit: financial resources, technological resources** **Shape: needs-based**	Value judgement in the need to redistribute resources in order to take on mitigation action in Namibia. The sentence implies that there is a basic level of resources that can be reached in which measures can be taken. This level is determined on the needs of the county in terms of financial and technological resources.
[F8] COP26 Philippines	Those who have polluted and continue to pollute the Earth's environment through unthinking industrialization starting 200 years ago must pay for the grants, investments, and subsidies needed for the most vulnerable countries to adapt to climate change.
Principle: Prioritarian **Topic: other** **Unit: financial resources** **Shape: proportional to contribution**	Prescribing the moral obligation to help the most vulnerable countries with climate change adaptation. Motivated by the idea that it is morally just to help them, no link to the aim of reducing inequalities is presented, making it prioritarian. A preferred shape of the distribution is proportional to historic contributions, with historically polluting countries having a responsibility to provide a larger share of financial resources.
[F9] COP20 Kenya	The INDCs must reflect the efforts Parties are willing to contribute towards the enhanced implementation of their obligations under the convention as enshrined in its Article 4.
Principle: egalitarian **Topic: UNFCCC agreements and principles** **Unit: Responsibility** **Shape: Proportional to commitments**	Moral judgement on what INDCs (Intended Nationally Determined Contributions) should at least include. Article 4 calls on the need for fair and transparent reporting on countries climate efforts as well as making a distinction between the responsibilities of on the implementation of commitments made by all in article 4, calling for the reporting of contributions by all .

Appendix E | Ground Truth Dataset - Categories

E1 | Bar plots of category labels

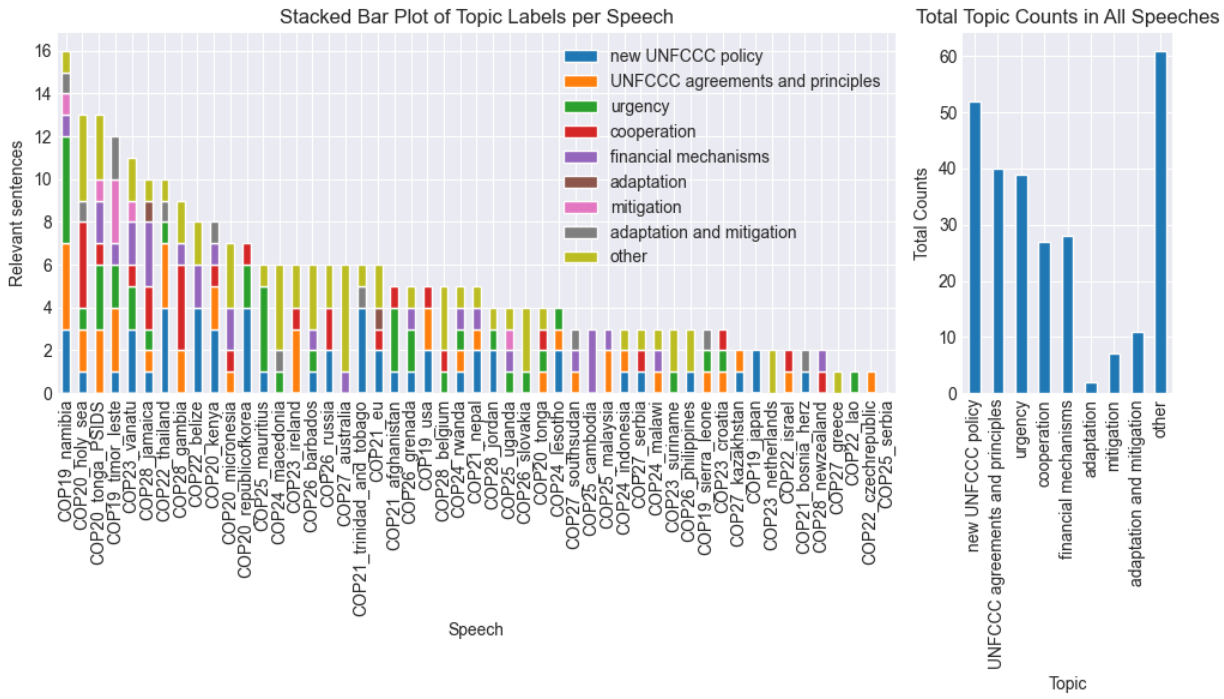

Figure 2: Left: Bar plot of Topic labels per speech - right: label counts over all speeches, indicating class imbalances

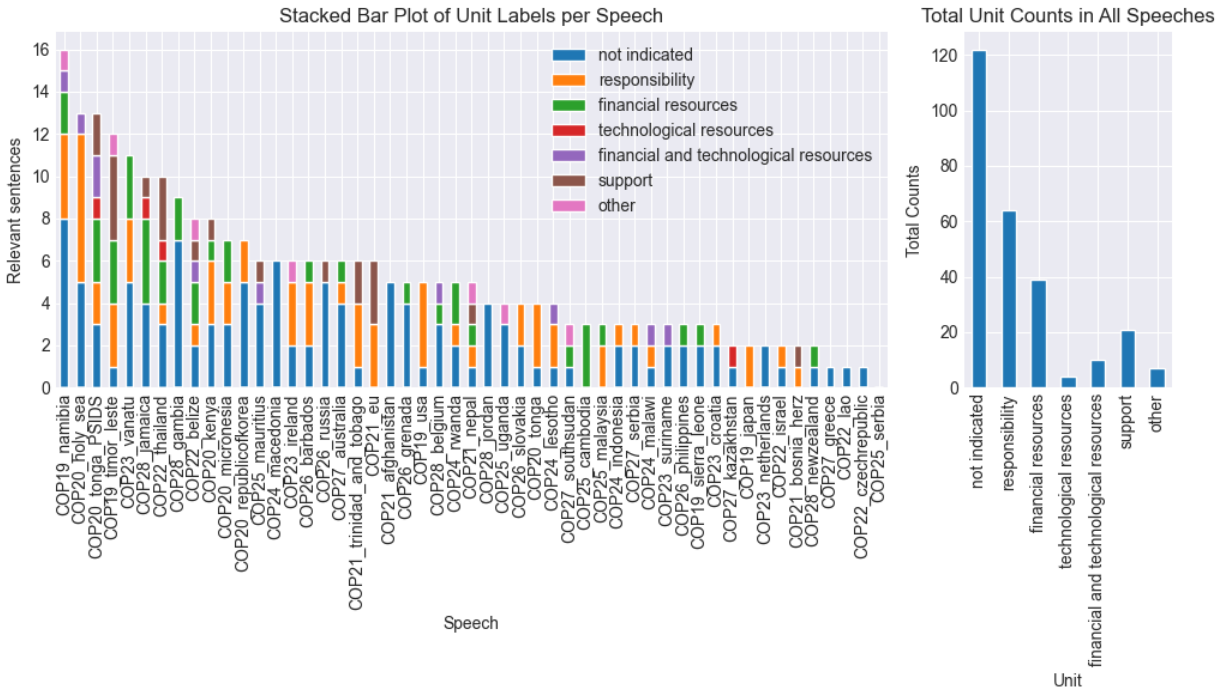

Figure 6: Left: Bar plot of Unit labels per speech - right: label counts over all speeches, indicating class imbalances

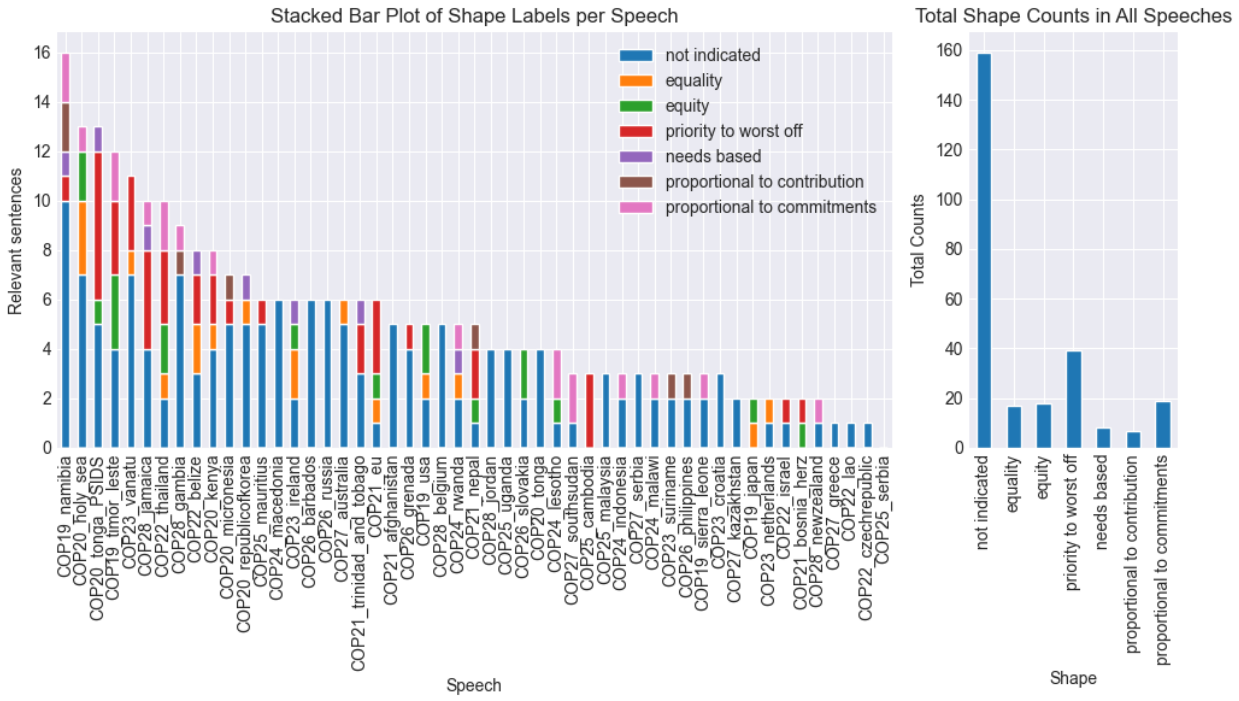

Figure 7: Left: Bar plot of Shape labels per speech - right: label counts over all speeches, indicating class imbalances

E2 | Correlation heatmap of categories and principles

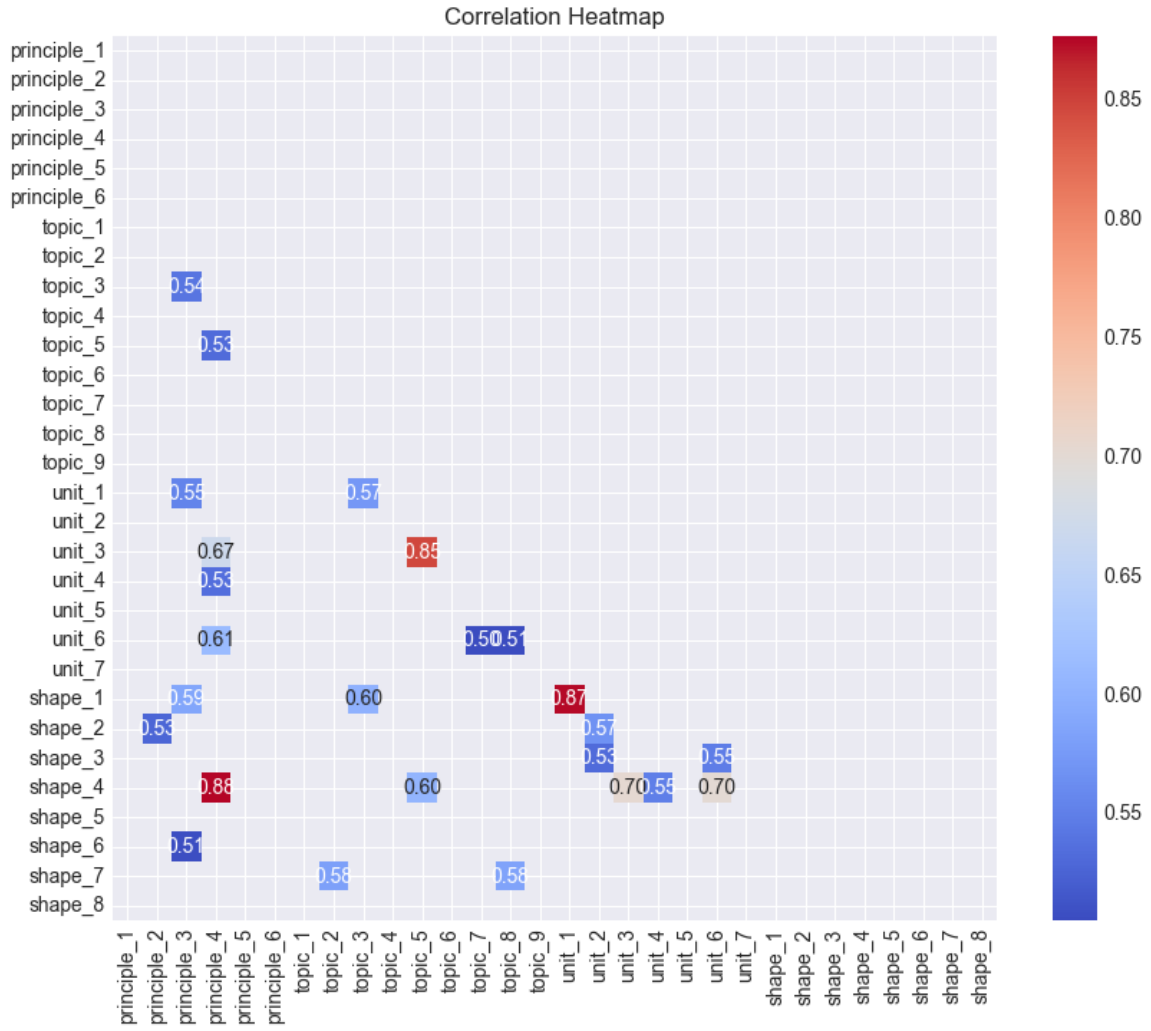

Figure 8 Correlation heatmap for all principles, topics, units, and shapes identified in manual annotation. Visualisation threshold is set at 0.5.

PRINCIPLE
1 general normative statement
2 egalitarian
3 utilitarian
4 prioritarian
5 sufficientarian
6 libertarian

TOPIC
1 new UNFCCC policy
2 UNFCCC agreements and principles
3 urgency
4 cooperation
5 financial mechanisms
6 adaptation
7 mitigation
8 adaptation and mitigation
9 other

UNIT
1 not indicated
2 responsibility
3 financial resources
4 technological resources
5 financial and technological resources
6 support
7 other

SHAPE
1 not indicated
2 equality
3 equity
4 priority to worst off
5 needs based
6 proportional to contribution
7 proportional to commitments

Appendix F | Test and Train set

Table 25 presents an overview of the class representation between the test and train dataset. Although an 80/20 split has been made, class level balances do not follow this distribution. This is because the split is made on speech level rather than sentence level. This was necessary to enable inclusion of contextual sentences during annotation.

Table 25 Appendix E - Class balances of the test and train set. Speeches included in the test set are: COP19 USA, COP20 Micronesia, COP21 Nepal, COP22 Belize, COP23 Ireland, COP24 Rwanda, COP25 Serbia, COP26 Barbados, COP27 Kazakhstan, COP28 Belgium

Category	Label	Train	Test	Total	Train as % of Total	Test as% of Total
RELEVANCE	Not relevant	786	213	999	79%	21%
	Statement of intent	208	69	277	75%	25%
	Relevant	218	49	267	82%	18%
PRINCIPLE	*Not evaluated*	992	281	1273	78%	22%
	General normative statement	33	17	50	66%	34%
	Egalitarian	51	16	67	76%	24%
	Utilitarian	59	6	65	91%	9%
	Prioritarian	66	6	72	92%	8%
	Sufficientarian	8	2	10	80%	20%
	Libertarian	1	3	4	25%	75%
TOPIC	*Not evaluated*	992	281	1273	78%	22%
	New UNFCCC Policy	41	11	52	79%	21%
	UNFCCC agreements and principles	32	9	41	78%	22%
	Urgency	36	4	40	90%	10%
	Cooperation	23	4	27	85%	15%
	Financial mechanisms	22	7	29	76%	24%
	Adaptation	2	-	2	100%	0%
	Mitigation	7	-	7	100%	0%
	Adaptation and mitigation	11	-	11	100%	0%
	Other	46	15	61	75%	25%
UNIT	*Not evaluated*	992	281	1273	78%	22%
	Not indicated	106	18	124	85%	15%
	Other	4	3	7	57%	43%
	Responsibility	49	15	64	77%	23%
	Financial resources	31	9	40	78%	23%
	Financial and technological resources	8	2	10	80%	20%
	Technological resources	3	1	4	75%	25%
	Support	19	2	21	90%	10%
SHAPE	*Not evaluated*	992	281	1273	78%	22%
	Not indicated	130	29	159	82%	18%
	Priority to worst off	34	5	39	87%	13%
	Proportional to commitment	19	1	20	95%	5%
	Proportional to contribution	5	2	7	71%	29%
	Equity	14	4	18	78%	22%
	Equality	11	6	17	65%	35%
	Needs based	5	3	8	63%	38%

Appendix G | Prompts

Orange: Notion to take context into account *Bx.x.1*
Green: Examples to take into account *Bx.1*

Prompt B3 and B4, an additional label is created: not evaluated – this label indicates sentences that are not seen as relevant or as statements of intent. .

G1 | Prompt B1

"You are tasked to annotate sentences from speeches presented at the High Level Segment at UNFCCC Conference of the Parties.
Label for the category named: ['RELEVANCE']. You must take the previous and next two sentences into account as context when determining the code you assign. You choose from the following codes: [Not relevant, Statement of intent, Relevant]

CODE: Not relevant; Description: A sentence that presents a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; Example: "We need adequate, predictable, accessible and sustainable finance."
CODE: Statement of intent; Description: A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification; Example: "Mr. President, as we come to the end of the International Year of the SIDS, we hope to leave Lima on a high note."
CODE: Relevant; Description: A sentence that presents a normative statement with a value judgment or prescription based on norms and values, including motivations derived from the textual context; Example: “I am sure that you all agree with me that we must collectively and responsibly act for the welfare of our common Planet - Terra – a tiny beautiful little blue dot on the Celestial Map of our Solar System which supports Life as we know it.”

Present your output in the following format:
unique_id | RELEVANCE

Annotate the following sentences based on the instructions provided:"

G2 | Prompt B2

"You are tasked to annotate sentences from speeches presented at the High Level Segment at UNFCCC Conference of the Parties. It is your task to identify by what Distributive Justice principle is foundational to the normative statements.
Label for the category named: ['PRINCIPLE']. The category name is: PRINCIPLE

You choose from the following codes: [general normative statement, egalitarian, utilitarian, prioritarian, sufficientarian, libertarian].
CODE: general normative statement; Description: A sentence that presents no indication of foundational distributive justice principle that is used as a motivation for the normative statement; Example: "There must be a common framework for regular reporting and tracking so that we know where we are headed; in addition to processes for reviewing and enhancing commitments, as well as for ensuring accountability and compliance."
CODE: egalitarian; Description: A sentence presenting an egalitarian motivation. Motivated by the goal to reduce inequalities and have equality of opportunity. This includes statements the highlight the inclusion of all and references to mutual contributions by all. The principle of common but differentiated responsibilities is also seen as egalitarian; Example: Example: "It is critical to develop common fair and transparent rules for carbon pricing measurement and ensure their multilateral recognition and comparability between countries."
CODE: utilitarian; Description: A sentence presenting a utilitarian motivation. Motivated by the goal of maximizing the benefit of all or contributing to 'the greater good'. The benefit of all can be in present and future generations. Found in sentences prescribing the need to take action, motivated by the need to improve the lives of all; Example: "We must activate all our efforts to preserve the environment for future generations, and related to it I would like to thank to all those who, despite the many obstacles and difficulties, continue in these efforts."
CODE: prioritarian; Description: A sentence presenting a prioritarian motivation. Motivated by the moral obligation to help the worst off. Highlighted in sentences focussing on supporting vulnerable developing countries. It does not focus existing differences; Example: "Furthermore, where adaptation is not an option, countries must have support for irreparable loss and damage"
CODE: sufficientarian; Description: A sentence presenting a sufficientarian motivation. Motivated by the idea that everyone should be able to reach a common level of well-being. Found in sentences referring to a global standard of well-being that everyone should reach as well as statements implying that no one should be left behind this threshold; Example: "The longer we wait, the more it will cost; more victims will suffer from our inaction and the greatest weight will fall on the most vulnerable, the poorest peoples and future generations: what is at issue here is respect for their fundamental human rights."
CODE: libertarian; Description: A sentence presenting a libertarian motivation. Motivated by a need for freedom, indicating minimal government involvement and focus on free markets; Example: "We should make the most of these opportunities by developing and implementing innovative solutions, enabling our businesses to contribute to, and benefit from, the global transition towards a carbon - neutral economy."

Label for the category named: ['PRINCIPLE'].
Your output should be in the following format:
unique_id | PRINCIPLE

Annotate the following sentences based on the instructions provided:"

G3 | Prompt B3

"You are tasked to annotate sentences from speeches presented at the High Level Segment at UNFCCC Conference of the Parties. It is your task to identify by what Distributive Justice principle is foundational to the normative statements.
Label for the category named: ['PRINCIPLE']. The category name is: PRINCIPLE. You must take the previous and next two sentences into account as context when determining the code you assign.

You choose from the following codes: [general normative statement, egalitarian, utilitarian, prioritarian, sufficientarian, libertarian].
CODE: not evaluated; Description: A sentence that does not present a normative statement, presenting a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; OR A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification; Example: "We need adequate, predictable, accessible and sustainable finance."
CODE: general normative statement; Description: A sentence that presents no indication of foundational distributive justice principle that is used as a motivation for the normative statement; Example: "There must be a common framework for regular reporting and tracking so that we know where we are headed; in addition to processes for reviewing and enhancing commitments, as well as for ensuring accountability and compliance."
CODE: egalitarian; Description: A sentence presenting an egalitarian motivation. Motivated by the goal to reduce inequalities and have equality of opportunity. This includes statements the highlight the inclusion of all and references to mutual contributions by all. The principle of common but differentiated responsibilities is also seen as egalitarian; Example: "It is critical to develop common fair and transparent rules for carbon pricing measurement and ensure their multilateral recognition and comparability between countries."
CODE: utilitarian; Description: A sentence presenting a utilitarian motivation. Motivated by the goal of maximizing the benefit of all or contributing to 'the greater good'. The benefit of all can be in present and future generations. Found in sentences prescribing the need to take action, motivated by the need to improve the lives of all; Example: "We must activate all our efforts to preserve the environment for future generations, and related to it I would like to thank to all those who, despite the many obstacles and difficulties, continue in these efforts."
CODE: prioritarian; Description: A sentence presenting a prioritarian motivation. Motivated by the moral obligation to help the worst off. Highlighted in sentences focussing on supporting vulnerable developing countries. It does not focus existing differences; Example: "Furthermore, where adaptation is not an option, countries must have support for irreparable loss and damage"
CODE: sufficientarian; Description: A sentence presenting a sufficientarian motivation. Motivated by the idea that everyone should be able to reach a common level of well-being. Found in sentences referring to a global standard of well-being that everyone should reach as well as statements implying that no one should be left behind this threshold; Example: "The longer we wait, the more it will cost; more victims will suffer from our inaction and the greatest weight will fall on the most vulnerable, the poorest peoples and future generations: what is at issue here is respect for their fundamental human rights."
CODE: libertarian; Description: A sentence presenting a libertarian motivation. Motivated by a need for freedom, indicating minimal government involvement and focus on free markets. Example: "We should make the most of these opportunities by developing and implementing innovative solutions, enabling our businesses to contribute to, and benefit from, the global transition towards a carbon - neutral economy."

Your output should be in the following format:
unique_id | PRINCIPLE

Annotate the following sentences based on the instructions provided:"

G4 | Prompt B4

For evaluation of the relevant train dataset, the red elements are removed.

"You are tasked to annotate sentences from speeches presented at the High Level Segment at UNFCCC Conference of the Parties. It is your task to identify by what Distributive Justice principle is foundational to the normative statements.
Label for the categories named: ['PRINCIPLE', 'TOPIC', 'UNIT', 'SHAPE']. The category names are: PRINCIPLE, TOPIC, UNIT, SHAPE

You choose from the following codes:
PRINCIPLE: [not evaluated, general normative statement, egalitarian, utilitarian, prioritarian, sufficientarian, libertarian]
TOPIC: [not evaluated, new UNFCCC policy, UNFCCC agreements and principles, urgency, cooperation, financial mechanisms, adaptation, mitigation, adaptation and mitigation, other]
UNIT: [not evaluated, not indicated, responsibility, financial resources, technological resources, financial and technological resources, support, other]
SHAPE: [not evaluated, not indicated, equality, equity, priority to worst off, needs based, proportional to contribution, proportional to commitment]

PRINCIPLE CODE: not evaluated; Description: A sentence that does not present a normative statement, presenting a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; OR A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.
PRINCIPLE CODE: general normative statement; Description: A sentence that presents no indication of foundational distributive justice principle that is used as a motivation for the normative statement.
PRINCIPLE CODE: egalitarian; Description: A sentence presenting an egalitarian motivation. Motivated by the goal to reduce inequalities and have equality of opportunity. This includes statements the highlight the inclusion of all and references to mutual contributions by all. The principle of common but differentiated responsibilities is also seen as egalitarian.
PRINCIPLE CODE: utilitarian; Description: A sentence presenting a utilitarian motivation. Motivated by the goal of maximizing the benefit of all or contributing to ‘the greater good’. The benefit of all can be in present and future generations. Found in sentences prescribing the need to take action, motivated by the need to improve the lives of all.
PRINCIPLE CODE: prioritarian; Description: A sentence presenting a prioritarian motivation. Motivated by the moral obligation to help the worst off. Highlighted in sentences focussing on supporting vulnerable developing countries. It does not focus existing differences.
PRINCIPLE CODE: sufficientarian; Description: A sentence presenting a sufficientarian motivation. Motivated by the idea that everyone should be able to reach a common level of well-being. Found in sentences referring to a global standard of well-being that everyone should reach as well as statements implying that no one should be left behind this threshold.
PRINCIPLE CODE: libertarian; Description: A sentence presenting a libertarian motivation. label principle_6 as 1 if presenting a libertarian motivation. Motivated by a need for freedom, indicating minimal government involvement and focus on free markets.

TOPIC CODE: not evaluated; Description: A sentence that does not present a normative statement, presenting a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; OR A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.

TOPIC CODE: new UNFCCC policy; Description: A sentence that discusses requirements for new UNFCCC policy
TOPIC CODE: UNFCCC agreements and principles; Description: A sentence that highlight articles, agreements or principles of the UNFCCC
TOPIC CODE: urgency; Description: A sentence that discusses the need to take action to address the climate crisis and implement new policies.
TOPIC CODE: cooperation; Description: A sentence that focuses on the need for cooperation.
TOPIC CODE: financial mechanisms; Description: A sentence that highlights financial redistributions based on a financial mechanism, for example the Green Climate Fund.
TOPIC CODE: adaptation; Description: A sentence that focuses on adaptation, adaptation measures, or adaptation policies.
TOPIC CODE: mitigation; Description: A sentence that focuses on mitigation, mitigation measures or mitigation policies.
TOPIC CODE: adaptation and mitigation; Description: A sentences that highlight both elements of adaptation and mitigation.
TOPIC CODE: other; Description: A sentence that discusses a topic that is not covered by any other TOPIC code.

UNIT CODE: not evaluated; Description: A sentence that does not present a normative statement, presenting a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; OR A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.
UNIT CODE: not indicated; Description: A sentence that does not highlight a specific unit of distribution.
UNIT CODE: responsibility; Description: A sentence that highlights the distribution of the responsibility to take on action in the context of climate change and climate measures.
UNIT CODE: financial resources; Description: A sentence that addresses the distribution of financial resources or money.
UNIT CODE: technological resources; Description: A sentence that addresses the distribution of technological resources.
UNIT CODE: financial and technological resources; Description: A sentence that highlights a distribution of both technological and financial resources.
UNIT CODE: support; Description: A sentence that indicates a distribution of support, without specifying what type of support.
UNIT CODE: other; Description: A sentence that discusses a distribution with a unit that is not covered by any other UNIT code.

SHAPE CODE: not evaluated; Description: A sentence that does not present a normative statement, presenting a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions; OR A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.
SHAPE CODE: not indicated; Description: A sentence that does not highlight a specific distributional shape.
SHAPE CODE: equality; Description: A sentence that highlights a distributional shape that is equal for all. The size of the allocation does not have to be specified. Sentences highlighting the need for “everyone to take action”, indicating that equality is favoured in the distribution of the obligation to take measures.
SHAPE CODE: equity; Description: A sentence that highlights a distributional shape that accounts for differences between the recipients, e.g. accounting for the financial situation of recipients.

SHAPE CODE: priority to worst off; Description: A sentence that prescribes a distribution only focussing on distribution to the ones that are deemed to be the worst off, e.g. the distribution of financial resources to countries that are most vulnerable to climate impacts.
SHAPE CODE: needs based; Description: A sentence prescribing a distribution that is based on the needs of recipients or what is deemed needed.
SHAPE CODE: proportional to contribution; Description: A sentence that indicates a distribution that is based on historical contributions. e.g. in relation to historical responsibility for past emissions, where historically higher polluters have to reduce more emissions.
SHAPE CODE: proportional to commitment; Description: A sentence indicating that the distribution should follow the shape of commitments made in global agreements, pledges, and treaties. e.g. found in statements calling upon developed countries to act upon their pledges to commit to the green climate fund.

Your output should be in the following format:
unique_id | PRINCIPLE | TOPIC | UNIT | SHAPE

Annotate the following sentences based on the instructions provided:"

G5 | Numerical prompt

Codebook used for numerical annotation of relevance.

"You are tasked to annotate sentences from speeches presented at the High Level Segment at UNFCCC Conference of the Parties.
Label for the categories named: ['relevance_0', 'relevance_1', 'relevance_2'].

You assign a binary label [1,0]: choose 1 if the category applies to the sentence, choose 0 if the category does not apply to the sentence.

CATEGORY: relevance_0; Description: A sentence that presents a descriptive or factual statement, including descriptions of policies already implemented or expressions of gratitude, condolences, formalities, or rhetorical questions.
CATEGORY: relevance_1; Description: A sentence with ambiguous normativity, often referring to policy intentions or expectations without clear normative motivations. These sentences prescribe an action or behaviour but lack explicit normative justification.
CATEGORY: relevance_2; Description: A sentence that presents a normative statement with a value judgment or prescription based on norms and values, including motivations derived from the textual context.

Present your output in the following format:
unique_id | relevance_0 | relevance_1 | relevance_2

Annotate the following sentences based on the instructions provided:"