Efficient Aspect-Based Summarization with Small Language Models: A Use-Case on Climate Change Reports

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have revolutionized many fields of Natural Language Processing (NLP), including summarization. These systems, however, consist of billions of parameters and, as such, they have the crucial shortcoming of being energy-intensive. In this work, we present a thorough evaluation of very recent, small-sized LLMs (SLMs) on the task of Aspect-Based Summarization of Climate Change Reports. In doing so, we show that modern SLMs are sufficiently good for the task and can bring value in assisting with summarization for policymakers while being more efficient than their bigger counterparts without significant performance deterioration. We also show how energy consumption among SLMs themselves does not correlate with better performance, further proving the point that smaller models can be effectively used for the task. Finally, we release the new dataset that we collected to perform our experiments, from which we hope research in NLP for climate change and research in efficient Aspect-Based Summarization with LLMs can develop further.

1 Introduction

Aspect-Based Summarization (ABS) is a popular task in Natural Language Processing (NLP), dealing with summarizing a text with respect to a specific aspect or topic (Titov and McDonald, 2008).

Recently, the landscape of NLP has seen a revolution happening in the form of Large Language Models (LLMs), which are capable of performing the majority of tasks that were previously performed by specifically trained systems, often outperforming the latter without the need for any supervision (Ziyu et al., 2023). These models, however, comprise billions of parameters, and, as such, their carbon footprint is one of the main factors leading to criticisms of their use in various areas in which smaller, comparable models are available (Faiz et al., 2024). These observations, as well as hardware constraints, have led to the development of smaller LLMs, which, notwithstanding the still comparatively higher number of parameters compared to previous systems, have been labeled as Small Language Models (SLMs) (Ranaldi and Freitas, 2024).

In this work, then we combine the latest development in SLMs with the task of ABS and we perform the first comprehensive evaluation in our knowledge of SLMs for the task. We do so by introducing a new domain for ABS, namely the one of climate change reports for which we introduce a new dataset. Climate change reports, in fact, are critical for policy-makers and researchers in tackling climatic challenges and, as such, fine-grained automatic summarization of such reports is a task in line with recent work advocating for ways in which NLP can help climate scientists and policymakers (Stede and Patz, 2021). Furthermore, the task itself is a natural benchmark for advocating the use of low-carbon LLMs.

The main questions informing our work are:

Q1: are SLMs comparable in performance to larger LLMs for our task?

Q2: among SLMs, is energy consumption positively correlated to performance on the task?

Q3: how do our models' performance deteriorate in the absence of ground truth paragraphs to summarize?

Our main contributions then are multiple:

1) We evaluate SLMs in the context of ABS.

2) We introduce a new dataset for the new domain of climate change reports within the scope of the task.

3) We focus on energy efficiency and we adapt an existing framework for energy-aware summarization evaluation to our use case while analyzing the correlation between energy consumption and performance. We present the first energy-aware comparison of modern LLMs for summarization and paving the way for future research in this area.

2 Related Work

2.1 Aspect-Based Summarization

ABS is the task of summarizing a given text with respect to a specific aspect or topic (Titov and McDonald, 2008). The task is particularly useful in aiding the reading of complex, multi-topic content such as news bulletins (Fremann and Klementiev, 2019) or Wikipedia articles (Hayashi et al., 2021).

In the context of ABS, the models developed for the task falls broadly in the category of supervised (Tan et al., 2020; Ma et al., 2022; Ahuja et al., 2022) and unsupervised models (Soleimani et al., 2022; Coavoux et al., 2019), where the firsts have shown improvements over the latter, but do need a sufficient number of training samples, for which there is a scarcity of data, especially in certain domains (Yang et al., 2023). More recently, modern LLMs have shown performance on par with previous supervised models also in unsupervised (i.e. zero-shot) setting for various NLP tasks (Ziyu et al., 2023) including summarization (Zhang et al., 2024). Such models are mostly under-explored in the context of ABS, as just isolated examples of their use for the task exist in the literature, which does not present comparisons between LLMs and SLMs and is limited to hotel reviews summarization (Jeong and Lee, 2024; Bhaskar et al., 2023).

In our work, then, we aim to fill this gap, while focusing on the efficiency and on the more specific domain of climate reports ABS.

2.2 SLMs and Efficiency Evaluation

Modern LLMs are extremely effective for a variety of tasks, but they comprise billions of parameters, leading to consideration of efficiency and environmental externalities associated with their use (Tokayev, 2023). These concerns have led to consider the overall environmental cost of such models when deploying them (Faiz et al., 2024).

At the same time, in the last year much effort has been spent in making the LLM landscape more efficient (Wan et al., 2024), either by proposing SLMs, yielding comparable results to LLMs thanks to refined datasets and knowledge distillation (Abdin et al., 2024; Team et al., 2024; Gu et al., 2024), or by exploring different types of quantization which can diminish the computational burden while maintaining a good trade-off with performance (Yao et al., 2024) or both.

Recent literature has proposed to include models' efficiency in evaluating summarization (Moro et al., 2023), but without including LLMs in their experiments. Much NLP literature has often ignored considerations about model efficiency, but as the models get bigger and the marginal improvements get smaller, including model efficiency in the evaluation is important for more sustainable and, ultimately, more usable NLP systems.

In this work, then, we draw also on literature on SLMs and efficiency evaluation in developing our experiments and then assessing them.

2.3 NLP and Climate Change

NLP can help with a variety of problems related to Climate Change including but not limited to: climate stance detection (Fraile-Hernandez and Peñas, 2024), climate-related question answering (Vaghefi et al., 2023; Biester et al., 2022) and automatic fact-checking (Meddeb et al., 2022; Mazid and Zarnaz, 2022). NLP can also improve access to information, which can be used for educational or policy-making purposes (Stede and Patz, 2021).

Our contribution, then, points in this direction and it builds on previous work to assess a new task in the area, namely that of ABS. Previous work, in fact, has drawn from data similar to the one we use in order to create a chatbot that can answer questions related to climate change with access to the most up-to-date information (Vaghefi et al., 2023). As new reports and new knowledge get produced at a fast pace, however, the need to assess the zero-shot ability of LLMs to summarize such reports in an efficient and fine-grained way is crucial to further help their reading from both policy-makers and researchers. No existing work in this direction exists in our knowledge and our work aims to fill this gap.

3 Methodology

3.1 Zero-Shot Aspect-Based Summarization with LLMs

In order to perform ABS with out-of-the-box LLMs and SLMs, we developed a simple prompt template which is presented to each model for a fair comparison. The prompt template T has the following format:

T="Summarize the main takeaways from the following text with respect to topic {topic}. Text: {text}"

We define the substitution function sub, which takes as inputs the template T, $topic$ and $text$ and substitutes {topic} and {text} in T with $topic$ and

$text$, respectively, thus obtaining:

$$prompt = sub(T, topic, text) \tag{1}$$

As we will see below, at times more than one paragraph needs to be summarized. Defining the collection of paragraphs to be summarized $P = \{p_1, ..., p_n\}$, where p_i are the individual paragraphs, we obtain:

$$text = \begin{cases} P, |P| = 1 \\ concat(P), |P| > 1 \end{cases} \tag{2}$$

where $concat$ indicates the concatenation of all the paragraphs in P.

The generation process, then, is done as:

$$\hat{y} = LLM(prompt) \tag{3}$$

Where LLM is the LLM currently used and $\hat{y}$ is the generated summary.

In many cases, there is also a limitation in the number of maximum tokens that some of the models can accept and especially in the case of many paragraphs p to be summarized the length of the input text might exceed this limit. Given this limitation, we also set a character threshold over which we get a set of interim results y^p_{int}:

$$y^p_{int} = LLM(sub(T, topic, p)) \forall p \in P \tag{4}$$

Then, having the collection Y_{int} of all y^p_{int}, we get the final text as:

$$text = concat(Y_{int}) \tag{5}$$

which can then be passed in equation 1 to obtain the final prompt to be passed in equation 3. The implications on the performance of such cases are further analyzed below.

3.2 Retrieval Augmented Generation

To answer Q3 and test the limits of our approach, we also investigate Retrieval Augmented Generation (RAG), where we automatically retrieve the k most relevant paragraphs from the given climate report and we use them as input for the LLM, instead of the ground truth paragraphs. This setting relates to the real-world use case in which, e.g., a policymaker wants an automatic system to both find the relevant information in the report and summarize it. Formally, we define an encoder model enc such that it encodes all the reports' paragraphs p_i as:

$$e_i = enc(p_i), e_i \in \Re^d \tag{6}$$

with d being the dimensionality of the embeddings from the given encoder enc. At inference time, the given aspect or topic $topic$ is encoded in the same embeddings space as:

$$q = enc(topic), q \in \Re^d \tag{7}$$

At this point, we define a number k of paragraphs that we want to retrieve from the collection of all paragraph indices $P_{ind} = \{1, ..., N\}$ and we retrieve the subset of paragraph indices $P_{sub} \subset P$ as:

$$P_{sub} = argmax_{i \in P_{ind}}(cos(q, e_i)), s.t. |P_{sub}| = k \tag{8}$$

where cos represents the cosine distance between the query embedding q and the given paragraph embedding e_i.

Having obtained the paragraphs associated with their indices in P_{sub}, we then obtain $text$ as described in equation 2. The final summary $\hat{y}$ is then obtained as:

$$\hat{y} = LLM(prompt_{rag}) \tag{9}$$

where $prompt_{rag}$ is obtained either with equation 1 or with equations 4 and 5 according to whether $text$ is longer than the character threshold as explained above.

3.3 Extractive Summarization Baseline

To compare the performance of LLMs with a non-generative baseline, we develop a simple extractive approach, based on the understanding of the task as a question-answering task. For each example, we again define an encoder enc and we follow equation 7 to obtain a query embedding q. Having obtained $text$ in one of the ways previously defined, we then divide it into sentences with the method by Kiss and Strunk (2006) and group them as $S = \{s_1, ..., s_n\}$ with n being the number of sentences in $text$. Each sentence s_i is then encoded as:

$$e^i_s = enc(s_i), e^i_s \in \Re^d \tag{10}$$

We define a number k of sentences to be extracted and the collection of all sentence indices in the document $S_{ind} = \{1, ..., n\}$ and we obtain its subset $S_{sub} \subset S_{ind}$ as:

$$S_{sub} = argmax_{i \in S_{ind}}(cos(q, e^i_s)), s.t. |S_{sub}| = k \tag{11}$$

The final summary is obtained by concatenating the sentences associated with such indices, that is:

$$\hat{y} = concat(s_i) \forall i \in S_{sub} \tag{12}$$

3.4 Evaluation

3.4.1 Aspect-Based Summarization Evaluation

Following recent research in the field of summarization evaluation, we use the ChatGPT-RTS (Shen et al., 2023) for evaluation. This metric uses the powerful ChatGPT LLM (i.e. GPT 3.5) as an evaluator, by framing the evaluation task as a question concerning the property of the summaries with respect to 4 key attributes individuated by Hayashi et al. (2021): coherence, consistency, fluency, and relevance. For each reference summary, paragraphs, and topic triplet, ChatGPT is given the definition of the dimension to evaluate as well as the triplet and asked to output a score from 1 to 5, together with an explanation for such a decision. We introduced a key modification to the relevance definition in the prompt to include the target topic so that, with minimal modification, the final score also takes into consideration the target aspect. In appendix A we illustrate in more detail the prompts fed to ChatGPT for performing the evaluation, as well as the correlation with human judgment and comparison with other metrics.

3.4.2 Retrieval Evaluation

To assess how successful different encoders are in retrieving the correct paragraphs in the RAG setting, we use the Mean Reciprocal Rank (MRR) metric, an information retrieval metric that considers how high in a ranked list the retriever can place the correct item (in our case the correct paragraph) (Radev et al., 2002).

In our case, we set the hyperparameter of MRR to 10, meaning that we consider the first 10 items as scored by the retriever as the limit beyond which we consider the retriever to have failed (leading to MRR@10 equals 0).

3.5 Energy Consumption and Efficiency Re-Weighting

The Carburacy method was proposed to account for efficiency in summarization evaluation, by re-weighting the ROUGE metric for summarization with the cost for running the model $C = E * D$, where E is the cost of a single example measured as the kg of CO_2 emitted by summarization models and D is the dataset size (Moro et al., 2023). The re-weighting formula is then applied as:

$$\gamma = \frac{e^{log_{\alpha} R}}{1 + C * \beta} \quad (13)$$

with R being the effectiveness score (i.e. the initial summarization metric) and $\alpha = 1$ and $\beta = 100$ following the original work. The authors further divided the costs in inference and training costs, but in our unsupervised setting just the first applies.

In applying the Carburacy re-weighting scheme to our context we took into account the fact that LLMs can lead to very different outcomes in terms of summaries length and this has an effect on the cost C as longer sequences will lead to higher consumption in the auto-regressive setting of decoder-only modern LLMs. In our case, we want to isolate the cost of each LLM as a function solely of its architecture, rather than of its output. Therefore, we compute equation 13 by setting $D = 1$ and E such that:

$$E = Emission(LLM_{stop:k}(prompt_{fix})) \quad (14)$$

Where $prompt_{fix}$ is a fixed prompt for each system and $Emission$ is the function computing CO_2 emissions. The key of the above modification is represented by $LLM_{stop:k}$ which we define as a constrained generation from the given system, where the generation stops automatically at a token number k which we set to 10. This way, each LLM receive a prompt of same input and output a same-length output, and by keeping these factors constant we assure to measure just differences in emissions caused by structural differences between LLMs (e.g. number of parameters).

When applying Carburacy to the extractive baselines and to the RAG models, instead, we simply apply equation 13 with the cost of encoding $prompt_{fix}$ in the first case and with the cost of encoding the entire dataset D in the latter. In the retrieval experiments, we empirically set $\beta = 10000$ to account for the difference in emission scale.

We measure CO_2 levels with the codecarbon python library[1], leveraging CPU as well as GPU energy consumption.

4 Data

For the purpose of this work, we have collected and released the SumIPCC dataset, comprising 140 topic-annotated summaries and relative paragraphs from climate change reports. We used two reports from the authoritative Intergovernmental Panel on Climate Change (IPCC) as a data source. The reports we used are the synthesis reports AR5 (IPCC, 2014b) and AR6 (IPCC, 2023b) for two separate

[1] https://codecarbon.io/

Feature	AR5	AR6	All
Summaries	70	70	140
Paragraphs	34	38	72
Summary Topics	63	70	133
Summary Section Headers	4	3	7
Summary Sub-Section Headers	17	18	35
Paragraphs Section Headers	34	38	72

Table 1: Statistics of our IPCC-Sum dataset. For all features, we report the number of unique occurrences for the different subsets (AR5 and AR6), as well as for the whole dataset. It can be noticed how many topics are repeated in different summaries.

Model	Billions of Parameters	C
Qwen 0.5B	0.5	4.06e-05
Qwen 1.8B	1.8	4.19e-05
Qwen 4B	4	5.28e-05
Qwen 7B	7	5.63e-05
Gemma 2B	2	4.41e-05
Gemma 7B	7	6.41e-05
Phi 3	3.8	5.30e-05
Llama 3	8	6.20e-05
Mistral	7	6.03e-05
ChatGPT	$\sim$ 175	$\sim$ 3.86e-03
GPT4	$\sim$ 175	$\sim$ 3.86e-03
MPNet	0.11	1.65e-07

Table 2: Number of parameters and estimated energy cost C for the ABS models. In every case, we used the conventional abbreviated notation, e.g., e-05 to signify a multiplier of 10^{-5} for the given value. Model size does not perfectly correlate with energy consumption, as different architectures might have different efficiency.

years, 2014 and 2023, which collected the contributions of different working groups on a variety of topics related to climate change and linked policies. The two reports were chosen among the IPCC synthesis report collections as they both include accompanying publications named Summary for Policy-Makers (IPCC, 2014a, 2023a), which include short summaries related to specific topics and referring to paragraphs in the respective synthesis reports. Each summary includes the main highlights with regard to a specific topic as discussed in the report and it might refer to multiple paragraphs in the original report, in case the specific topic is treated in different parts of the report.

On occasions, we observed summaries that were too broad in scope, referring to many different long paragraphs, but comprising just a few lines on a broad topic: we filtered out these cases. The final result is a dataset comprising 140 paragraph-summary pairs with associated topics, which we manually annotated to be as precise as possible. Paragraphs and section headers from the Summary for Policy-Makers could also have been used to annotate the summaries, but they were ambiguous as they grouped different summaries; they are also included as features in the dataset, even though we don't explore their use in this work. As we will see, however, there are a number of summaries sharing the same topic but in different contexts and future work might include additional information to better disambiguate these cases, especially in the RAG context. Table 1 shows the features from the collected dataset and their occurrences, while Appendix E includes additional information.

5 Experimental Setup

5.1 LLMs and Extractive Baselines

We compare recent and popular LLMs: 9 open-source SLMs and 2 big, proprietary LLMs. For the SLMs, there is no single definition of how small a model should be to be considered such, therefore we impose a hardware constraint to choose the models, namely to be able to fit in a single NVIDIA® Tesla T4 GPU with 16GB of memory: to achieve this, we have then selected models up to 8 billion parameters, while using 4-bit quantization on all the models from this category; the effect of the quantization has been shown to be negligible in most cases (Yao et al., 2024). The SLMs we used are: Qwen 1.5 (Qwen) 0.5B, 1.8B, 4B and 7B (Bai et al., 2023), Gemma 1.1 (Gemma) 2B and 7B (Team et al., 2024), Phi 3 (Abdin et al., 2024), Llama 3 8B (Llama 3) (Meta, 2024) and Mistral v0.2 7B (Mistral) (Jiang et al., 2023). In every case, we have used the instruction-tuned versions of the models: we give additional details about the models' source and run time in Appendix F.

To compare the performance of SLMs with bigger LLMs, we compare them with the state-of-the-art GPT4 (OpenAI et al., 2024) and its earlier version, ChatGPT (Brown et al., 2020); no public information about the quantization settings nor the model size exist for the two models, but table 2 includes estimates on size and energy cost C for these models together with the actual models size and cost for the small-sized LLMs. We computed C as per equation 14, while we report a rough estimate of the sizes of GPT4 and ChatGPT by equating them to the size of the related model GPT3 (Brown et al., 2020) and we estimated their cost C by multiplying the cost of Gemma 2B for the module of the respective model parameters; this is indeed a very rough estimate, but it should give a good approximation of the scale of difference between

Model	Billions of Parameters	C
DistilRoB	0.08	4.06e-05
MPNet	0.11	4.19e-05
MiniLM	0.2	4.42e-10
GTR	1.2	5.63e-05
ST5	1.2	4.41e-05
GTE	0.44	6.41e-05

Table 3: Number of parameters and estimated energy cost C for the text encoders used as zero-shot retrievers in our RAG experiments.

Figure 1: Pearsons' correlation between the metrics' aspects and energy consumption.

small-sized LLMs and bigger, state-of-the-art ones. Finally, for the extractive baselines we have used the all-mpnet-base-v2 (MPNet) model, further described in the next section. Also for this models, we include the energy cost C in table 2.

5.2 Retrieval and Extractive Models

To choose the zero-shot text retrieval models for the RAG experiments, we have mostly drawn from the top open-source systems from the MTEB benchmarks evaluating out-of-the-box text embedding systems (Muennighoff et al., 2023). At the same time, we have included the same hardware constraints explained in section 5.1 to limit our choice to relatively small-sized encoders. The final models we used in the RAG setting, then, are: all-mpnet-base-v2 (MPNet), an encoder based on the MPNet architecture (Song et al., 2020) and on the sentence transformers framework (Reimers and Gurevych, 2019) to be highly performative in a variety of sentence encoding tasks, all-distilroberta-v1 (DistilRoB), a distilled version of RoBERTa (Liu et al., 2019) trained similarly to MPNet, all-MiniLM-L12-v2 (MiniLM), a small and extremely efficient transformer encoder (Gu et al., 2024) further fine-tuned similarly to MPNet, gtr-t5-xl (GTR) (Ni et al., 2022b) and sentence-t5-xl (ST5) (Ni et al., 2022a), two sentence encoders both based on the encoder part of the T5 architecture (Raffel et al., 2020) but fine-tuned on different datasets for text retrieval, and gte-large-en-v1.5 (GTE) (Li et al., 2023), a transformer encoder trained with multi-stage contrastive learning.

Table 3 shows the number of parameters for this set of models, together with the energy cost C computed as described in the methodology section.

6 Experiments

6.1 SLMs Evaluation

Table 5 shows the results obtained by running and comparing to reference summaries our SLMs and baselines over the SumIPCC dataset with the ground truth paragraphs for each reference summary (i.e. without RAG). The results clearly highlight a very good performance on behalf of most SLMs and LLMs, whereas the extractive baselines show inferior performance for all the given evaluation dimensions; such a difference is statistically significant ($p < 0.05$) and it confirms the superiority of LLMs of any size to the simple extractive models. It is interesting to notice, however, that the performance of the extractive method is generally good in absolute terms for the relevance and consistency dimensions, highlighting the style of this dataset, where many exact lines from the target paragraphs are present in the reference summaries (see appendix A and appendix E for more details).

When comparing SLMs with the LLMs baselines, we can observe some striking results in that the ChatGPT baseline appears to be the best-performing system overall, even more so than the superior GPT4 baseline. This apparently counter-intuitive result can, however, be explained by three factors: first, as the metric we use is based on ChatGPT itself it might show a bias in favor of the model, as observed in previous studies (Panickssery et al., 2024), second, the reliability of the metric in the context of high-quality summaries is generally lower (Shen et al., 2023), and third, ChatGPT is not significantly better than GPT4 in any evaluation dimension. These points also apply to most SLMs. More recent and relatively more powerful SLMs like Llama 3, in fact, appear to be worse than other models like ChatGPT itself, but ultimately the difference is statistically insignificant, rather indicating that most SLMs and LLMs perform similarly in our context. SLMs, then, can be as effective as larger LLMs for our task (Q1).

Turning to Q2, figure 1 shows how the energy consumption shows a weak, but relevant correlation with LLMs performance on each dimension. A key driver of this correlation is the poor performance of Qwen 0.5B, suggesting that there is a threshold under which model size severely impacts the capacity of SLMs to perform this task. The

Model	Consistency ↑	Coherence ↑	Fluency ↑	Relevance ↑	Average ↑
Qwen 0.5B	4.52*	4.33*	4.41*	4.06*	4.33*
Qwen 1.8B	4.89	4.83	4.88	4.79	4.85
Qwen 4B	4.75*	4.84	4.91	4.56*	4.77
Qwen 7B	4.84	4.94	4.9	4.74	4.86
Gemma 2B	4.86	4.86	4.96	4.71	4.85
Gemma 7B	4.85	4.94	4.99	**4.81**	4.9
Phi 3	4.84	4.92	4.94	4.74	4.86
Llama 3	4.82	4.84	4.91	4.74	4.83
Mistral	4.78*	4.84	4.95	4.6	4.79
ChatGPT	**4.94**	**4.96**	**4.98**	4.79	**4.91**
GPT4	4.83	4.89	4.96	**4.81**	4.89
MPNet	4.44*	3.03*	3.45*	4.15*	3.77*

Table 4: Summarization results for all dimensions obtained by evaluating our models with the ChatGPT-RTS metric. Asterisks indicate that the results are significantly worse than the best model (i.e. ChatGPT).

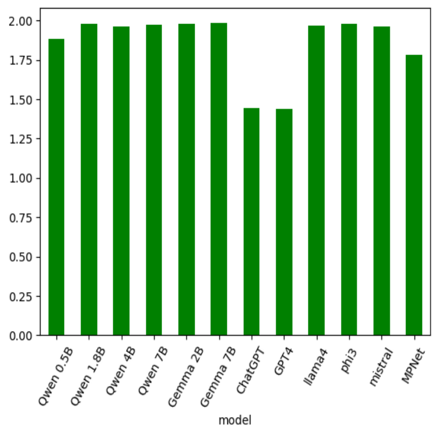

Figure 2: ChatGPT RTS Average scores re-weighted via Carburacy.

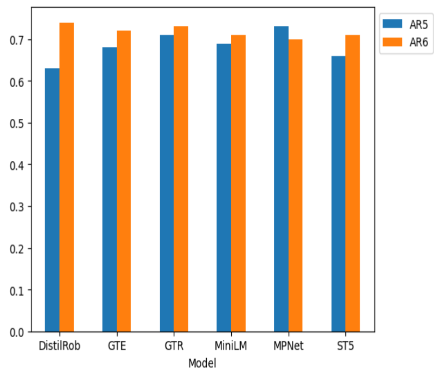

Figure 3: Retrieval results in terms of MRR@10 metric re-weighted via the Carburacy method.

updated ranking of models in figure 2 using the Carburacy technique, however, shows how on certain occasions, notably that of Qwen 1.8B, very small SLMs can perform similarly to larger ones. The re-ranking confirms once more that most SLMs perform similarly, and that are generally better than very small LLMs (Qwen 0.5B) and then the extractive baseline. It follows, that ChatGPT and GPT4 are actually the worst models when considering the efficiency/effectiveness trade-off because the increase in energy consumption is not justified by a relevant increase in the models' performance.

6.2 RAG Evaluation

Figure 3 shows the results of using different retrieval models on the two subsets of our dataset, separately. It can be seen how also in this case most models perform similarly and, applying the Carburacy method to re-weight the MRR@10 score, this leads to comparatively smaller models being the best choice to perform the retrieval in our context.

Having identified the best retrieval models for both subsets of our dataset, we employ them to retrieve the top 2 documents for each query and then we employ the method described in section 3.2 to generate the summaries. In this case, we have used just the best models for each family, as indicated by results in table 5. It is interesting to notice how this time the results from different models are more spread, highlighting more significant differences individuated by our metric in this more challenging scenario. This is in line with what was previously observed for the same metric, as using ChatGPT to evaluate ABS has been shown to be more accurate and more confident about its own decision when the difference in the quality of the generated summaries is substantial (Shen et al., 2023). The fact of using two paragraphs that might not be the correct ones as input to be summarized according to a specific topic, in fact, seems to have an effect on all dimensions, not only on the relevance one (which presents the biggest overall drop in performance, as it could have been expected). This evidence suggests that our task in a RAG setting is indeed a more challenging task, which requires further inves-

Model	Consistency ↑	Coherence ↑	Fluency ↑	Relevance ↑	Average ↑
Qwen 1.8B	3.66	**4.36**	4.24	3.11	3.84
Gemma 2B	3.21*	3.81*	3.67*	3.21	3.48*
Phi 3	3.32*	3.82*	3.74*	3.23	3.53*
Llama 3	**3.76**	4.27	**4.44**	**3.26**	**3.93**
Mistral	3.02*	3.61*	3.56*	3.02	3.30*
ChatGPT	3.24*	3.81*	3.52*	2.96	3.38*
MPNet	2.68*	2.39*	2.5*	2.36*	2.48*

Table 5: Summarization results for all dimensions obtained by evaluating our models with the ChatGPT-RTS metric on the retrieved passages. Asterisks indicate results that are significantly worse than the best model (i.e. Llama 3).

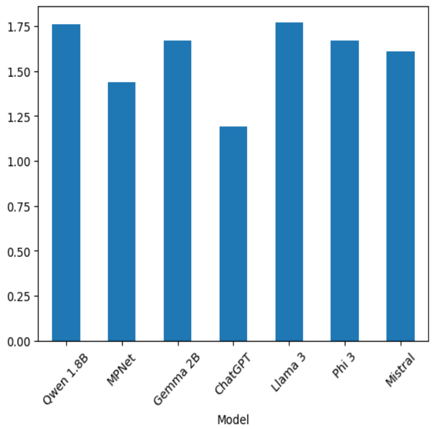

Figure 4: ChatGPT RTS Average scores for the RAG experiment re-weighted via the Carburacy method.

tigation both in terms of the retrieval model being used and in terms of the summarization model. Different LLMs, in fact, appear to be more capable of dealing with heterogeneous information and filter out irrelevant information, while maintaining good coherence, fluency, and consistency with the input paragraphs (more qualitative examples under this respect are presented in D). Because of this, in this context the choice of the model appears to be relevant, with Llama 3 performing significantly better than most other models, in line with the models' performance on existing benchmarks (Meta, 2024). Interestingly it can be seen how the much smaller Qwen 1.8B, however, performs similarly to Llama 3 and this leads to the model being ranked as good as the latter in the re-weighted results using Carburacy, shown in figure 4. This last evidence shows once more that smaller LLMs can perform as well as bigger ones in our context and this might be because of a variety of reasons including but not limited to training data, stochasticity, and prompt preferences: in deciding which model is best for a specific task, then, the inclusion of efficiency in the evaluation framework allows to identify models with a smaller energy-cost, while leading to a drop in performance which is minimal or even not significant.

7 Conclusion and Future Directions

In this work, we have investigated the use of SLMs for ABS in the context of climate change reports. Apart from the task itself, which has a variety of uses in policy-making and education, our aim was that of evaluate whether smaller, more efficient LLMs (i.e. SLMs) can lead to comparable results to bigger one in a task in which LLMs are extremely capable. The results indeed confirmed that SLMs are a valid alternative to bigger LLMs, especially in the easier scenario in which ground truth paragraphs were provided. As this task was easy enough to be solved by most LLMs, in fact, results were not significantly different in most cases, and applying a re-weighting scheme that takes into consideration the CO_2 emissions of the models helped identify the best model both in terms of efficiency and performance.

When we turned to the RAG scenario, instead, it could be seen that the difference in the models we used appeared to be more significant. Also in this case, however, the smallest model performed comparably with the best-performing one and, even though this might be due to various things not reflecting a more general equivalence, the evidence suggests, at least, that smaller models can be a valid alternative also in more challenging cases.

Finally, we release our dataset and this can lead to many interesting research directions both in terms of NLP applications for climate science and in terms of SLMs evaluation. Specifically, future research could explore the RAG setting further by incorporating more fine-grained information during retrieval (e.g. section and/or paragraph titles, which are included in the dataset) and fine-tune SLMs on the small available data to test the ability of such models to learn from small data. We leave these directions open for future research.

8 Limitations

Our work deals with the use of SLMs for ABS and has shown that they often perform similarly to larger LLMs in our context. Given the specific domain of application (i.e. climate change reports), however, we are limited to a small size dataset, which in turn increases results' variability. Another limitation of our work involve the evaluation metric, which includes a number of problems such as having around 80% agreement with human judgement, as shown in appendix A: this value is relatively high for summarization metrics, but it is still low enough to represent a significant limitation in terms of how much we can trust the metric itself in certain cases. Other evaluation limitations include the fact that our metric has been shown to correlate less with human judgement when dealing with high-performing systems (which is our case in the first experiment using ground truth paragraphs) and the already noticed fact that the metric appears to be biased towards certain LLMs (i.e. ChatGPT).

Finally, there is initial evidence that the aspects we have evaluated for each sample in our dataset might be too broad leading to the summarizers reporting redundant information in the summaries. Future research might consider using the additional features we provided in the released dataset in order to better define the aspect on which the summarization models should focus.

9 Ethical Considerations

Using LLMs and SLMs to summarize climate change reports raises several ethical considerations:

1) **Accuracy and Reliability**. If inaccurate or misleading summaries are produced by LLMs, this could potentially misinform stakeholders and the public, leading to poor decision-making. Therefore, it is essential to have a human-in-the-loop approach in double checking the summaries produced by such systems.

2) **Transparency and Accountability**. LLMs are black-box and therefore are not transparent nor accountable in terms of what output they produce. Notwithstanding the de-biasing and alignment with human preferences that the systems we used undertook, the reasons why such models produced certain summaries remain opaque.

3) **Representation Issues and Bias**. LLMs have been shown to include a number of biases derived from the training data. In the context of climate change reports, dealing with different world regions and cultures, this might lead to inaccurate and/or biased depiction of different populations.

4) **Accessibility and Inclusivity**. The use of LLMs require access to resources that might not be widely available in less developed countries and poorly funded institutions and, therefore, these could lead to problem of inclusivity and reduced access to our tool for policy-makers and educators from such background.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Laura Biester, Dorottya Demszky, Zhijing Jin, Mrinmaya Sachan, Joel Tetreault, Steven Wilson, Lu Xiao, and Jieyu Zhao, editors. 2022. *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Ahmad Faiz, Sotaro Kaneda, Ruhan Wang, Rita Osi, Prateek Sharma, , and Lei Jiang. 2024. LLMCarbon: modeling the end-to-end carbon footprint of large language models. In *International Conference on Learning Representations (ICLR)*.

Jesus M. Fraile-Hernandez and Anselmo Peñas. 2024. HAMiSoN-generative at ClimateActivism 2024: Stance detection using generative large language models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 79–84, St. Julians, Malta. Association for Computational Linguistics.

Lea Frermann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. *Preprint*, arXiv:2306.08543.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. WikiAsp: A Dataset for Multi-domain Aspect-based Summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

IPCC. 2014a. Climate change 2014: Summary for policy-makers. Technical report, IPCC.

IPCC. 2014b. Climate change 2014: Synthesis report. Technical report, IPCC.

IPCC. 2023a. Climate change 2023: Summary for policy-makers. Technical report, IPCC.

IPCC. 2023b. Climate change 2023: Synthesis report. Technical report, IPCC.

Nayoung Jeong and Jihwan Lee. 2024. An aspect-based review analysis using chatgpt for the exploration of hotel service failures. *Sustainability*, 16(4).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Tibor Kiss and Jan Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32:485–525.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Tinghuai Ma, Qian Pan, Huan Rong, Yurong Qian, Yuan Tian, and Najla Al-Nabhan. 2022. T-bertsum: Topic-aware text summarization based on bert. *IEEE Transactions on Computational Social Systems*, 9(3):879–890.

Md Abdullah Al Mazid and Zaima Zarnaz. 2022. Climate change myths detection using dynamically weighted ensemble based stance classifier. In *Proceedings of the 2nd International Conference on Computing Advancements*, ICCA '22, page 277–283, New York, NY, USA. Association for Computing Machinery.

Paul Meddeb, Stefan Ruseti, Mihai Dascalu, Simina-Maria Terian, and Sebastien Travadel. 2022. Counteracting french fake news on climate change using language models. *Sustainability*, 14(18).

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Gianluca Moro, Luca Ragazzi, and Lorenzo Valgimigli. 2023. Carburacy: Summarization models tuning and comparison in eco-sustainable regimes with a novel carbon-aware accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14417–14425.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022a. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022b. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Dragomir R. Radev, Hong Qi, Harris Wu, and Weiguo Fan. 2002. Evaluating web-based question answering systems. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Leonardo Ranaldi and Andre Freitas. 2024. Aligning large and small language models via chain-of-thought reasoning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Amir Soleimani, Vassilina Nikoulina, Benoit Favre, and Salah Ait Mokhtar. 2022. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 49–62, Dublin, Ireland. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Kassym-Jomart Tokayev. 2023. Ethical implications of large language models a multidimensional exploration of societal, economic, and technical concerns. *International Journal of Social Analytics*, 8(9):17–33.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Commun. Earth Environ.*, 4(1).

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Preprint*, arXiv:2312.03863.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Zhewei Yao, Xiaoxia Wu, Cheng Li, Stephen Youn, and Yuxiong He. 2024. Exploring post-training quantization in llms from comprehensive study to low rank compensation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19377–19385.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.

A Metric Correlation with Human Judgement

Previous research has variously shown how summarization metrics are generally unreliable, yielding low correlation with human judgement; the use of ChatGPT in this context was observed to be the method yielding results more similar to the judgement expressed by human annotators, with correlation values around 0.50 (Shen et al., 2023). Still, our use case was slightly different from the one in the above work, as it deals with ABS rather than normal summarization and, given the specificity of our dataset (see appendix A) it also includes various snippets of texts directly copied from the main text in the reference summaries.

To assess the reliability of different metrics in this context and to choose which to report, we have asked two human annotators to rank 10 pairs of summaries generated by different LLMs and then we compared the results thus obtained with the ranking produced by different summarization metrics. Table 6 shows the results thus obtained in terms of percentage of matches between human annotators' rankings and the metrics obtained by recent metrics based on LLMs. It can be seen how ChatGPT RTS far outperforms the alternatives reaching very high agreement with the human annotators (close to 80%).

If we consider the agreement with traditional, similarity-based metrics depicted in figure 5, we can also observe how the the majority of traditional metrics generally agree with human annotators in this task at a level close to the one reached by ChatGPT RTS. This is indeed quite specific to the dataset we are considering as summaries are often presented as highlights reporting entire sentences from the source paragraph and, as LLMs are asked to generate highlights as well, rather than summaries, similarity-based metrics are actually quite good in this scenario. As traditional metrics lack a distinction between different dimensions of the generated summaries, however, we opted for ChatGPT RTS as the metric for our main experiments, as it yields similar agreement with human annotators, but with the added value of giving a multi-dimensional evaluation.

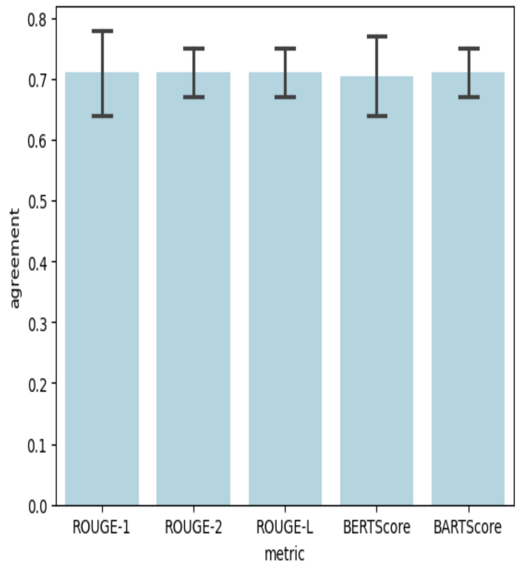

Figure 5: Average percentage of agreement between human annotators and similarity-based summarization metrics: standard deviation is also included in the form of error bars.

B Evaluation Prompts

In using the ChatGPT RTS, we have prompted ChatGPT with 4 different prompts per summary, to evaluate the different dimensions of the generated summaries. For what concerns consistency, coherence and fluency, we have adopted the same prompts from Shen et al. (2023). For what concerns relevance, we re-adapted the original formulation to make it fit for ABS, where we want our summary to be relevant with respect to a specific topic, in addition to the reference summary, where the original formulation did not include any topic nor reference summary.

We refer the reader to the original formulation in Shen et al. (2023) for the prompt used for consistency, coherence and fluency dimensions. For the relevance dimension, we show the prompt we used in figure 6.

Metric	Consistency	Coherence	Fluency	Relevance
ChatGPT RTS	0.77 ± 0.0	0.83 ± 0.06	0.66 ± 0.11	0.77 ± 0.0
ChatGPT MCQ	0.06 ± 0.06	0.55 ± 0.0	0.17 ± 0.06	0.44 ± 0.0
UniEval	55 ± 0.11	0.61 ± 0.06	0.33 ± 0.22	0.67 ± 0.0

Table 6: Average percentage of agreement between human annotators and LLM-based summarization metrics: standard deviation is also included.

```
Choose an option from A to E in order to score the following Aspect-Based
Summary given the corresponding Article, Aspect
    and Reference Summary with respect to relevance
    from one to five, where one indicates "irrelevant" and five indicates "perfect
relevance".
    Note that relevance measures selection of important content from the
source which align with the given Aspect and Reference Summary. The
summary should include
    only important information related to the given Aspect from the source
document.

    Article: {article}

    Aspect: {aspect}

    Reference Summary: {reference_summary}

    Aspect-Based Summary: {summary}

    A: The Summary is totally irrelevant with the Article and Aspect.
    Score: One.
    B: The majority of the Summary is irrelevant with the
    Article and Aspect. Score: Two.
    C: Some information in the Summary is relevant with the
    Article and Aspect whereas some are not. Score: Three.
    D: The majority of the Summary is relevant with the
    Article and Aspect. Score: Four.
    E: All information included in the Summary is relevant
    with the Article and Aspect. Score: Five.
    Your Answer (enter 1 letter from A to E):
```

Figure 6: The prompt used for evaluation with ChatGPT with the ChatGPT RTS evaluation method for the relevance aspect. At inference time {article} is substituted with the target paragraphs, {aspect} is substituted with the aspect on which the summarizer should focus, {reference_summary} is replaced with the reference summary and {summary} is replaced with the generated summary. All other dimensions have been evaluated with similar prompts, but without the need of {reference_summary} and {aspect} and substituting the description of the dimension with the relevant description from the other dimensions, as described in Shen et al. (2023).

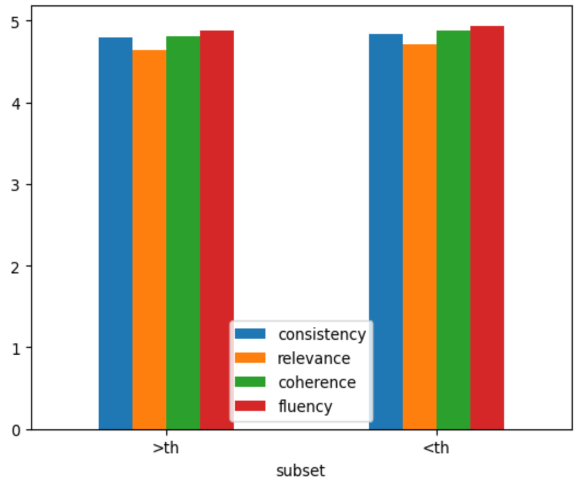

Figure 7: Comparison of performance in terms of ChatGPT RTS for instances longer (left) and shorter (right) than our fixed threshold (th).

C Effect of Long Inputs

In the methodology section, we highlighted how when using SLMs for summarization is usual to find instances in which input paragraphs are longer than the allowed token limits for the model. We have tackled these instances by applying an iterative procedure where we summarize individual paragraphs and then we ask the given LLM to summarize the concatenation of the summaries (see 4). In order to ensure that such a process won't lead to drop in performance we have plotted the performance of instances in which the paragraphs are longer than the fixed character threshold that we have empirically set in our experiments and we compared them to the plotted performance of instances where paragraphs are shorter than the threshold. Figure 7 shows such comparison: as it can be noticed, the difference in performance shows longer inputs performing marginally worse due to the iterative procedure. Still, the difference is minimal suggesting that our method does not negatively affect performance and, as such, it is a feasible way to overcome models' context limits.

Prompt:

Summarize the main takeaways from the following text with respect to the topic: Technological Solutions
Text: Technology development and transfer Technology policy complements other mitigation policies across all scales from international to sub-national, but worldwide investment in research in support of GHG mitigation is small relative to overall public research spending. Technology policy includes technology-push and demand-pull. Such policies address a pervasive market failure because, in the absence of government policy such as patent protection, the invention of new technologies and practices from R&D efforts has aspects of a public good and thus tends to be under-provided by market forces alone. Technology support policies have promoted substantial innovation and diffusion of new technologies, but the cost-effectiveness of such policies is often difficult to assess. Technology policy can increase incentives for participation and compliance with international cooperative efforts, particularly in the long run. Many adaptation efforts also critically rely on diffusion and transfer of technologies and management practices, but their effective use depends on a suitable institutional, regulatory, social and cultural context. Adaptation technologies are often familiar and already applied elsewhere. However, the success of technology transfer may involve not only the provision of finance and information, but also strengthening of policy and regulatory environments and capacities to absorb, employ and improve technologies appropriate to local circumstances.

ChatGPT

The main takeaways from the text regarding the topic of technological solutions are:

1. Investment in research for greenhouse gas (GHG) mitigation is relatively low compared to overall public research spending.
2. Technology policy, which includes both technology-push and demand-pull approaches, is necessary to address market failures and promote the development and adoption of new technologies.
3. Technology support policies have shown success in promoting innovation and diffusion of new technologies; however, assessing their cost-effectiveness is challenging.
4. Technology policy can incentivize participation and compliance in international cooperative efforts in the long term.
5. Adaptation efforts also rely on technology transfer, but their effectiveness depends on suitable institutional, regulatory, social, and cultural contexts.
6. Technology transfer involves not only providing finance and information, but also strengthening policy and regulatory environments and building capacities to absorb and improve technologies according to local circumstances.

Llama 3

Here are the main takeaways with respect to the topic of Technological Solutions:

1. **Technology policy complements other mitigation policies**: Technology development and transfer are essential components of a comprehensive approach to mitigate greenhouse gas emissions.
2. **Market failure**: The invention of new technologies and practices from R&D efforts is a public good, which tends to be under-provided by market forces alone, highlighting the need for government policy support.
3. **Technology-push and demand-pull**: Technology policy includes both technology-push (encouraging innovation) and demand-pull (stimulating adoption) approaches to promote the development and diffusion of new technologies.
4. **Cost-effectiveness assessment challenges**: The effectiveness of technology support policies is often difficult to assess, making it essential to monitor and evaluate their impact.
5. **Adaptation efforts rely on technology transfer**: Adaptation efforts rely on the diffusion and transfer of technologies and management practices, which requires a suitable institutional, regulatory, social, and cultural context.
6. **Success depends on local context**: The success of technology transfer depends not only on the provision of finance and information but also on strengthening policy and regulatory environments and capacities to absorb, employ, and improve technologies appropriate to local circumstances.

Qwen 0.5B

The text provides several takeaways about technological solutions:

1. Technology policy complements other mitigation policies across all scales.
2. Technology support policies promote substantial innovation and diffusion of new technologies.
3. Technology policy can increase incentives for participation and compliance with international cooperation efforts.
4. Adaptation technologies are often familiar and already applied elsewhere, but the success of technology transfer may involve not only the provision of finance and information, but also strengthening of policy and regulatory environments and capacities to absorb, employ, and improve technologies appropriate to local circumstances.

Overall, these takeaways suggest that technological solutions are essential for mitigating global climate change and promoting sustainable economic growth.

Reference

Technology policy complements other mitigation policies across all scales, from international to sub-national; many adaptation efforts also critically rely on diffusion and transfer of technologies and management practices. Policies exist to address market failures in R&D, but the effective use of technologies can also depend on capacities to adopt technologies appropriate to local circumstances.

Figure 8: An example of three summaries obtained using three LLMs: ChatGPT, Llama 3 and Qwen 0.5B. Prompt indicates the command given to the LLMs, including the text to be summarized and the target aspect. Reference indicates the reference human-generated summary. It can be seen how all models, even the smaller Qwen 0.5B, manage to produce sensible summaries, even though they do include extra information with respect to the reference summary (for which a more specific aspect formulation might be needed).

D Qualitative Evaluation

Figure 8 show an example of summaries generated for a given reference by different LLMs, together with the reference summary and the prompt used to obtain the summaries, including the target ground truth paragraph to be summarized. When ground truth target paragraphs are included, it can be seen that all LLMs give sensible answers which are comparable to each other. Some redundant information is included in all cases, but specifying the aspect more strictly is likely to solve that problem. When retrieved paragraphs are used in the RAG setting, instead, LLMs struggle to produce sensible results, as the discrepancy between the input (incorrect) paragraph and the aspect to be summarized tend to confuse the models, as highlighted in figure 9: this effect is stronger for weaker models as evident from the significance of the results in table 5 and, looking picture 9, from the case of Mistral, which produced a summary which is relatively long and mostly unrelated to the target aspect.

Prompt

Summarize the main takeaways from the following text with respect to the topic: Technological Solutions
Text: Technology development and transfer Technology policy complements other mitigation policies across all scales from international to sub-national, but worldwide investment in research in support of GHG mitigation is small relative to overall public research spending. Technology policy includes technology-push and demand-pull. Such policies address a pervasive market failure because, in the absence of government policy such as patent protection, the invention of new technologies and practices from R&D efforts has aspects of a public good and thus tends to be under-provided by market forces alone. Technology support policies have promoted substantial innovation and diffusion of new technologies, but the cost-effectiveness of such policies is often difficult to assess. Technology policy can increase incentives for participation and compliance with international cooperative efforts, particularly in the long run. Many adaptation efforts also critically rely on diffusion and transfer of technologies and management practices, but their effective use depends on a suitable institutional, regulatory, social and cultural context. Adaptation technologies are often familiar and already applied elsewhere. However, the success of technology transfer may involve not only the provision of finance and information, but also strengthening of policy and regulatory environments and capacities to absorb, employ and improve technologies appropriate to local circumstances. Common enabling factors and constraints for adaptation and mitigation responses Adaptation and mitigation responses are underpinned by common enabling factors. These include effective institutions and governance, innovation and investments in environmentally sound technologies and infrastructure, sustainable livelihoods and behavioural and lifestyle choices. Innovation and investments in environmentally sound infrastructure and technologies can reduce greenhouse gas emissions and enhance resilience to climate change. Innovation and change can expand the availability and/ or effectiveness of adaptation and mitigation options. For example, investments in low-carbon and carbon-neutral energy technologies can reduce the energy intensity of economic development, the carbon intensity of energy, GHG emissions, and the long-term costs of mitigation. Similarly, new technologies and infrastructure can increase the resilience of human systems while reducing adverse impacts on natural systems. Investments in technology and infrastructure rely on an enabling policy environment, access to finance and technology and broader economic development that builds capacity. Adaptation and mitigation are constrained by the inertia of global and regional trends in economic development, GHG emissions, resource consumption, infrastructure and settlement patterns, institutional behaviour and technology. Such inertia may limit the capacity to reduce GHG emissions, remain below particular climate thresholds or avoid adverse impacts. Some constraints may be overcome through new technologies, financial resources, increased institutional effectiveness and governance or changes in social and cultural attitudes and behaviours. Vulnerability to climate change, GHG emissions, and the capacity for adaptation and mitigation are strongly influenced by livelihoods, lifestyles, behaviour and culture . Shifts toward more energy-intensive lifestyles can contribute to higher energy and resource consumption, driving greater energy production and GHG emissions and increasing mitigation costs. In contrast, emissions can be substantially lowered through changes in consumption patterns. The social acceptability and/or effectiveness of climate policies are influenced by the extent to which they incentivize or depend on regionally appropriate changes in lifestyles or behaviours. Similarly, livelihoods that depend on climate-sensitive sectors or resources may be particularly vulnerable to climate change and climate change policies. Economic development and urbanization of landscapes exposed to climate hazards may increase the exposure of human settlements and reduce the resilience of natural systems. For many regions and sectors, enhanced capacities to mitigate and adapt are part of the foundation essential for managing climate change risks. Such capacities are place- and context-specific and therefore there is no single approach for reducing risk that is appropriate across all settings. For example, developing nations with low income levels have the lowest financial, technological and institutional capacities to pursue low-carbon, climate-resilient development pathways. Although developed nations generally have greater relative capacity to manage the risks of climate change, such capacity does not necessarily translate into the implementation of adaptation and mitigation options. Improving institutions as well as enhancing coordination and cooperation in governance can help overcome regional constraints associated with mitigation, adaptation and disaster risk reduction. Despite the presence of a wide array of multilateral, national and sub-national institutions focused on adaptation and mitigation, global GHG emissions continue to increase and identified adaptation needs have not been adequately addressed. The implementation of effective adaptation and mitigation options may necessitate new institutions and institutional arrangements that span multiple scales .

ChatGPT

- Investment in research for GHG mitigation is small relative to overall public research spending.
- Technology policies can address market failure and promote innovation and diffusion of new technologies.
- The cost-effectiveness of technology support policies is often difficult to assess.
- Technology policy can increase incentives for participation and compliance with international efforts.
- The success of technology transfer depends on suitable institutional, regulatory, social, and cultural contexts.
- Adaptation and mitigation responses require effective institutions, innovation, investments, and sustainable livelihoods.
- Innovation and investments in environmentally sound infrastructure and technologies can reduce emissions and enhance resilience.
- Constraints to adaptation and mitigation include inertia in economic development, emissions, infrastructure, and technology.
- Vulnerability to climate change and mitigation capacity are influenced by livelihoods, behavior, and culture.
- Shifts towards more energy-intensive lifestyles contribute to higher energy consumption and GHG emissions.
- Adaptation and mitigation capacities are place-specific and there is no one-size-fits-all approach.
- Developing nations have lower capacities for pursuing low-carbon, climate-resilient development pathways.
- Enhancing institutions, coordination, and cooperation can help overcome regional constraints.
- New institutions and arrangements may be necessary for effective adaptation and mitigation.

Llama 3

The main takeaways from the text with respect to technological solutions are:

1. **Technology policy** is essential to promote innovation and diffusion of new technologies, but its cost-effectiveness can be difficult to assess.
2. **Technology transfer** is critical for adaptation efforts, but its success depends on a suitable institutional, regulatory, social, and cultural context.
3. **Innovation and investments** in environmentally sound technologies and infrastructure can reduce greenhouse gas emissions and enhance resilience to climate change.
4. **Enabling factors** for adaptation and mitigation responses include effective institutions and governance, innovation and investments in environmentally sound technologies and infrastructure, sustainable livelihoods, and behavioral and lifestyle choices.
5. **Constraints** to adaptation and mitigation responses include the inertia of global and regional trends in economic development, GHG emissions, resource consumption, infrastructure, and settlement patterns, as well as institutional behavior and technology.
6. **Vulnerability to climate change** is influenced by livelihoods, lifestyles, behavior, and culture, and can be reduced through changes in consumption patterns and lifestyles.
7. **Capacity building** is essential for developing nations with low income levels, which have limited financial, technological, and institutional capacities to pursue low-carbon, climate-resilient development pathways.
8. **Institutional arrangements** that span multiple scales are necessary to implement effective adaptation and mitigation options and overcome regional constraints associated with mitigation, adaptation, and disaster risk reduction.

Mistral

The text discusses the role of technology policy in addressing greenhouse gas (GHG) emissions and adapting to climate change. The text highlights that technology policy complements other mitigation policies and is essential due to market failures in the invention and diffusion of new technologies. The text also emphasizes the importance of innovation and investments in environmentally sound technologies and infrastructure to reduce GHG emissions and enhance resilience.

The text also mentions that adaptation efforts critically rely on technology transfer and the importance of a suitable institutional, regulatory, social, and cultural context for effective use. Common enabling factors for adaptation and mitigation include effective institutions and governance, innovation, and investments in environmentally sound technologies and infrastructure.

The text also mentions that vulnerability to climate change, GHG emissions, and the capacity for adaptation and mitigation are influenced by lifestyles, behavior, and culture. Shifts towards more energy-intensive lifestyles can contribute to higher energy and resource consumption, driving greater energy production and GHG emissions. In contrast, emissions can be substantially lowered through changes in consumption patterns.

The text also mentions that developing nations have the lowest financial, technological, and institutional capacities to pursue low-carbon, climate-resilient development pathways. Improving institutions and enhancing coordination and cooperation in governance can help overcome regional constraints associated with mitigation, adaptation, and disaster risk reduction.

In summary, the text emphasizes the importance of technology policy in addressing GHG emissions and adapting to climate change, the role of innovation and investments in environmentally sound technologies and infrastructure, and the influence of lifestyles, behavior, and culture on vulnerability and capacity for adaptation and mitigation. The text also highlights the importance of effective institutions and governance in overcoming regional constraints.

Figure 9: An example of three summaries obtained using three LLMs: ChatGPT, Llama 3 and Mistral (the weakest LLM among this set of experiments). Prompt indicates the command given to the LLMs, including the text to be summarized and the target aspect. The reference summary is depicted in figure 8.

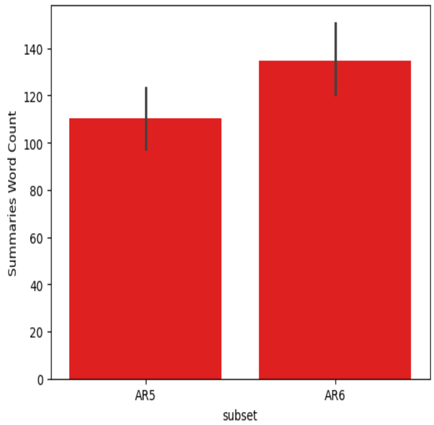

Figure 10: Average word count in the reference summaries for the two subsets of our dataset.

Figure 12: Most common summary topics in the AR5 subset of our dataset.

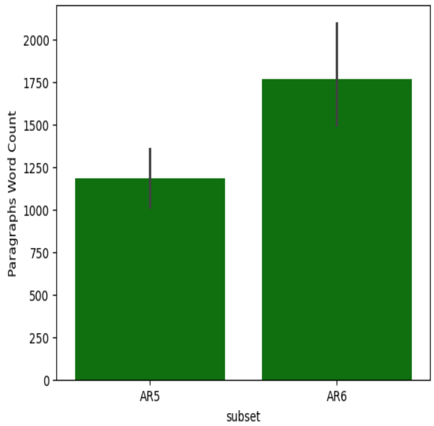

Figure 11: Average word count in the target paragraphs for the two subsets of our dataset.

Figure 13: Most common summary topics in the AR6 subset of our dataset.

E Dataset Statistics

Here, we present more in depth statistics for our SumIPCC dataset which we release under MIT license. Specifically, we report average word counts in summaries (figure 10) and in target paragraphs (figure 11), more common words in the summaries' topics for AR5 (figure 12) and AR6 (figure 13) subsets and lexical overlaps between reference summaries and target paragraphs in terms of rouge-1, rouge-2 and rouge-l (figure 14).

Overall, topics are similar between the two subsets and AR5 generally includes shorter paragraphs and shorter summaries than AR6. Also, it is evident by comparing figures 10 and 11 how the compression rate is quite high. Finally, figure 14 show how the lexical overlap between reference summaries and target paragraphs is also quite high reflecting the nature of the summaries often reflecting highlights rather than abstractive summaries.

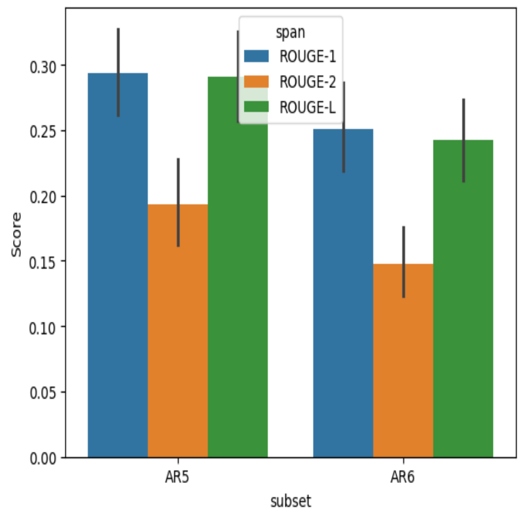

Figure 14: Rouge-1, rouge-2 and rouge-l scores of the reference summaries with respect to the target full paragraphs. These metrics represent the general overlap of the summaries with respect to the paragraphs, which is overall quite high in our case.

F Model Details

In our experiments we have used in all cases the pre-trained models as hosted on Huggingface Hub, but for ChatGPT and GPT4, for which we have used the official API.

Specifically, we report below the link for each of the open-source models we used:

1. Qwen 0.5B: https://huggingface.co/Qwen/Qwen1.5-0.5B-Chat
2. Qwen 1.8B: https://huggingface.co/Qwen/Qwen1.5-1.8B-Chat
3. Qwen 4B: https://huggingface.co/Qwen/Qwen1.5-4B-Chat
4. Qwen 7B: https://huggingface.co/Qwen/Qwen1.5-7B-Chat
5. Llama 3: https://huggingface.co/meta-llama/Meta-Llama-3-8B
6. Gemma 2B: https://huggingface.co/google/gemma-1.1-2b-it
7. Gemma 7B: https://huggingface.co/google/gemma-1.1-7b-it
8. Phi 3: https://huggingface.co/microsoft/Phi-3-mini-128k-instruct
9. Mistral: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

The models were all quantized in 4 bit with the bitandbytes python library[2] and run on a single NVIDIA® T4 GPU[3] with 16GB of RAM, as previously explained. All the models run between 2.5 and 10 hours, depending on model size and length of generated summaries: no sampling was applied for replicability.

Details of the GPT models we used are presented in table 7:

Model	Model Official Name	Revision
ChatGPT	gpt-35-turbo-16k	0613
GPT4	gpt-4	0125-Preview

Table 7: Details of the used GPT models.

Notice that throughout this work we have used the term ChatGPT to refer to GPT 3.5, consistently with previous literature (Shen et al., 2023): this naming is, however, erroneous as ChatGPT refers to the service rather than the underlying model.

[2]https://github.com/TimDettmers/bitsandbytes
[3]https://www.nvidia.com/en-us/data-center/tesla-t4/