

# Can Reasoning LLMs Synthesize Complex Climate Statements?

Yucheng Lu

New York University, New York, USA

yuchenglu@nyu.edu

## Abstract

Accurately synthesizing climate evidence into concise statements is crucial for policy making and fostering public trust in climate science. Recent advancements in Large Language Models (LLMs), particularly the emergence of reasoning-optimized variants, which excel at mathematical and logical tasks, present a promising yet untested opportunity for scientific evidence synthesis. We evaluate state-of-the-art reasoning LLMs on two key tasks: (1) *contextual confidence classification*, assigning appropriate confidence levels to climate statements based on evidence, and (2) *factual summarization of climate evidence*, generating concise summaries evaluated for coherence, faithfulness, and similarity to expert-written versions. Using a novel dataset of 612 structured examples constructed from the Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), we find reasoning LLMs outperform general-purpose models in confidence classification by 8 percentage points in accuracy and macro-F1 scores. However, for summarization tasks, performance differences between model types are mixed. Our findings demonstrate that reasoning LLMs show promise as auxiliary tools for confidence assessment in climate evidence synthesis, while highlighting significant limitations in their direct application to climate evidence summarization. This work establishes a foundation for future research on the targeted integration of LLMs into scientific assessment workflows. Code and data are publicly available at <https://github.com/YuchengLu-NYU/LLMClimateSynthesis>.

## 1 Introduction

Climate science involves complex systems, intricate modeling approaches, and specialized terminology that create significant barriers to public understanding (Stermann, 2011; Somerville and Hassol, 2011; Bernauer and McGrath, 2016). Despite overwhelming scientific consensus on climate

change, this complexity hinders widespread awareness and informed decision-making, even among policymakers responsible for addressing this global challenge (Pidgeon and Fischhoff, 2011). The extensive body of scientific evidence, while providing nuanced understanding of the systems and causal mechanisms driving climate change, simultaneously complicates efforts to communicate clear, actionable information—a fundamental challenge at the intersection of science, policy, and public engagement (van Eck, 2023). Large Language Models (LLMs) offer promising capabilities for addressing this communication gap. With their ability to process and synthesize vast amounts of text data, LLMs could potentially serve as powerful tools for distilling complex climate science into accessible formats (To et al., 2024; Bulian et al., 2024a). However, the nuanced nature of scientific evidence in climate research, with its inherent uncertainties and complex causal relationships, presents challenges that may exceed the capabilities of general-purpose LLMs (Bulian et al., 2024b). Recent developments in AI have produced specialized reasoning-optimized LLMs, which are explicitly designed to perform multi-step logical analysis and incorporate chain-of-thought processes that mirror analytical reasoning. These models are trained using reinforcement learning techniques to improve their ability to handle complex logical and mathematical problems (Cheng et al., 2025). In this study, we evaluate two state-of-the-art (SOTA) reasoning LLMs: DeepSeek-R1 (DeepSeek-AI et al., 2025) and OpenAI’s o3-mini (OpenAI, 2025).<sup>1</sup> As a baseline, we also perform the same two tasks on

<sup>1</sup>We selected o3-mini over OpenAI’s flagship reasoning model o1-pro and o1 due to availability and cost considerations. At the time of writing, o1-pro is not available as an API, whereas o1 costs \$60.00 per million output tokens, including reasoning tokens, compared to o3-mini’s \$4.40 and DeepSeek-R1’s regular price of \$2.19 (discount price \$0.55). These cost differences have significant implications for practical applications in research and deployment settings.

GPT-4o, one of the most widely used and capable general-purpose LLMs available. Our research makes three key contributions:

- We develop a focused dataset of 612 structured examples derived from the IPCC AR6, specifically designed for evaluating climate science evidence synthesis. Though modest in size, this curated resource offers high-quality pairs of scientific evidence bases with expert-written summaries and standardized confidence levels, providing a specialized benchmark for both classification and generative tasks in climate communication.
- To our knowledge, we conduct the first evaluation of reasoning LLMs for climate evidence synthesis, assessing their ability to assign appropriate confidence levels to climate statements based on presented evidence. Moreover, we show that the strong performance of LLMs is not the result of pure memorization by benchmarking against “no evidence” prompts, where we provide reference to specific sections in IPCC AR6 but withhold actual evidences in context.
- We evaluate these models’ summarization abilities on complex climate evidence, revealing important insights about the distinct skills required for effective scientific communication versus classification tasks. This analysis highlights the specific capabilities needed for translating scientific evidence into accessible formats for policymakers and the public.

These contributions collectively advance our understanding of how AI systems might address the critical challenge of communicating climate science more effectively, potentially facilitating greater public understanding and more informed policymaking in this crucial domain.

## 2 Related Work

**Climate Science and NLP** The application of NLP techniques to climate science has gained increasing popularity in recent years (Stammach et al., 2024). Incorporating artificial intelligence in the assessment and communication of climate statements is among the most important research directions within the Climate NLP research program. Costa et al. (2024) introduced ClimateQ&A,

a dataset and LLM-based assistant that answers climate and biodiversity-related questions grounded in scientific reports from the IPCC and IPBES, which builds upon previous related works (Morio and Manning, 2023; De-Gol et al., 2023; Muccione et al., 2024; Schimanski et al., 2024; Mullappilly et al., 2023).

However, research specifically focusing on climate evidence synthesis and assessment remains nascent. Joe et al. (2024) conducted a preliminary evaluation of GPT-4o’s capabilities for climate change evidence synthesis and systematic assessments, but primarily focused on information extraction rather than comprehensive evidence evaluation. Similarly, Li et al. (2024b) extracted climate change statements in IPCC reports to understand patterns of confidence levels and evidence types, while Lacombe et al. (2023) developed CLIMATEX, which assessed statements from IPCC AR6 reports without their supporting evidence bases. These works emphasized information retrieval capabilities of general-purpose LLMs rather than evidence synthesis or confidence attribution.

Our work differs significantly by evaluating models’ abilities to not only extract climate knowledge but to synthesize evidence and assign appropriate confidence levels—tasks more directly aligned with scientific communication needs. Furthermore, we specifically examine reasoning-optimized LLMs, which have not previously been evaluated for climate evidence synthesis tasks.

**Reasoning LLMs** Recent advancements in LLMs have led to specialized variants designed specifically for reasoning tasks. These models incorporate architectural innovations and targeted training methodologies to enhance their logical and multi-step reasoning capabilities. DeepSeek-R1 and OpenAI’s o3-mini represent SOTA examples in this class of models, balancing exceptional performance with computational efficiency.

The broader landscape of reasoning in LLMs has been extensively studied. Huang and Chang (2023) provides a comprehensive survey of reasoning capabilities in LLMs, identifying key methodologies that enable more sophisticated logical analysis. Notably, Wei et al. (2022) demonstrated that chain-of-thought prompting significantly enhances reasoning performance across various benchmarks. Both DeepSeek-R1 and OpenAI’s o3-mini incorporate explicit chain-of-thought in their reasoning. Additionally, Sun et al. (2024) categorizes

various reasoning frameworks in foundation models, emphasizing the unique strengths of models optimized for reasoning tasks. Xu et al. (2025) surveyed the application of reinforcement learning (RL) in improving LLMs’ reasoning capacity, a training technique employed by both DeepSeek-R1 and o3-mini.

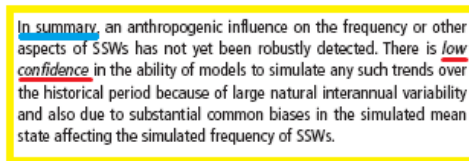
Despite these advances, the application of reasoning LLMs to scientific evidence synthesis remains relatively unexplored, particularly in domains like climate science where uncertainty quantification and nuanced interpretation are essential for effective communication and policy guidance.

**Evidence Synthesis with LLMs** The task of synthesizing scientific evidence and assigning appropriate confidence levels has traditionally been performed by human experts following established protocols (IPCC, 2010; Mastrandrea et al., 2011). Recent work by (Van Veen et al., 2023; Peng et al., 2023; Delgado-Chaves et al., 2025) explored the use of LLMs for evidence synthesis in medical contexts, finding promising capabilities while acknowledging significant challenges remain, especially regarding trust and robustness. However, evidence synthesis in the climate domain remains largely unexplored. Reasoning LLMs, with their enhanced capabilities for logical analysis, represent a particularly promising approach for addressing the unique challenges of climate evidence synthesis, where nuanced interpretation of evidence is essential for effective science communication and policy guidance.

### 3 Dataset

**The Intergovernmental Panel on Climate Change (IPCC) Sixth Assessment Report (AR6)** IPCC AR6 represents the most comprehensive synthesis of climate science to date, compiled by hundreds of leading scientists and approved by 195 member governments. Published between 2021 and 2023, AR6 consists of contributions from three Working Groups covering the physical science basis (IPCC AR6 WGI Masson-Delmotte et al. (2021)), impacts and adaptation (IPCC AR6 WGII Pörtner et al. (2022)), and mitigation of climate change (IPCC AR6 WGIII Shukla et al. (2022)), along with a Synthesis Report that integrates findings across all components. A distinguishing feature of the IPCC AR6 is its rigorously structured format that follows a systematic evidence-to-conclusion framework. Each section

presents detailed evidence bases drawn from peer-reviewed literature, followed by carefully crafted summary statements with explicitly assigned confidence levels. These confidence assessments follow a standardized methodology (Mastrandrea et al., 2011) that combines scientific agreement and evidence quality, producing calibrated language that expresses varying degrees of certainty (see Figure 5 in Appendix B for details). This structured approach makes AR6 an ideal source for systematic extraction of evidence-conclusion pairs with associated confidence assessments. Figure 1 illustrates



In summary, an anthropogenic influence on the frequency or other aspects of SSWs has not yet been robustly detected. There is low confidence in the ability of models to simulate any such trends over the historical period because of large natural interannual variability and also due to substantial common biases in the simulated mean state affecting the simulated frequency of SSWs.

Figure 1: Example conclusion from IPCC AR6 WGI

a sample conclusion from the the *Sudden Stratospheric Warming Activity* subsection from Chapter 3 *Human Influence on the Climate System* from IPCC AR6 WGI Masson-Delmotte et al. (2021). Figure 7 in the Appendix shows the subsection, which includes section header, evidence bases, and conclusion in its original layout.

The report’s consistent organization enables reliable parsing of the relationship between supporting evidence and resulting conclusions. Each finding is traceable to its underlying evidence base<sup>2</sup>, with transparent reasoning that connects specific climate observations, model outputs, and scientific theories to summary statements. This evidence-conclusion structure, combined with standardized confidence metrics, provides a gold-standard dataset for evaluating how effectively LLMs can process complex scientific information, determine appropriate confidence levels, and generate accurate summaries that preserve key scientific content while maintaining appropriate expressions of certainty.

**Data Extraction Process** We follow a three-step procedure to extract evidence-conclusion data pairs.

<sup>2</sup>Note that evidences presented in these subsections are already summaries with interpretations produced by climate experts, much like the exposition of literature in the related work or literature review sections of any scientific publication. That being said, for future research, one might be interested in retrieving the original, source research articles and having LLMs synthesizing from ground up.

1. **Document Preprocessing:** We converted PDF files to Markdown format using MinerU (Wang et al., 2024), a SOTA open-source PDF information extraction tool.<sup>3</sup> Given the extensive length of AR6 reports, we segmented them into manageable chunks based on the reports’ table of contents. We incorporated one-page overlaps between segments to prevent information loss at section boundaries, as often one section begins on the same page where the previous section ends.
2. **Argument Identification:** We parsed each Markdown file using header tags (#) to identify distinct sections. To ensure the extraction of genuine evidence-conclusion pairs, we applied filtering criteria to identify argumentative sections. A section was classified as containing an argument if it: (1) consisted of three or more paragraphs, and (2) concluded with a paragraph containing one of the following concluding phrases: “in summary”, “to summarize”, “in conclusion”, “overall,”<sup>4</sup>, “to conclude”, “in short”, or “to sum up”. While this approach may have excluded some valid evidence-conclusion pairs, it prioritized data quality over quantity.
3. **Confidence Level Extraction:** We identified and extracted the confidence levels associated with each conclusion. For conclusions containing multiple assessments with distinct confidence levels, we segmented the conclusion paragraph into individual statements. For example, the statement “To conclude, atmospheric aerosols sampled by ice cores, influenced by northern mid-latitude emissions, show positive trends from 1700 until the last quarter of the 20th century and decreases thereafter (*high* confidence), but there is *low* confidence in observations of systematic changes in other parts of the world in these periods” was divided into two separate conclusions with their respective confidence levels. Since there are too few “very low” and “very high” confidence conclusions at the end of the process, we keep only conclusions with

<sup>3</sup>MinerU allows the extraction of pictures. However, we choose to disregard these pictures for the sake of fairness in comparison. While GPT-4o allows pictures as inputs, reasoning LLM APIs do not currently take pictures as input.

<sup>4</sup>The “,” comma after overall is important to reduce false positives.

“low”, “medium”, or “high” confidence.

We deliberately employed a rule-based parsing strategy rather than relying on LLMs for data extraction to avoid potential issues of content hallucination or misrepresentation. Previous research by (Huang et al., 2023; Mohamed et al., 2025) has demonstrated that LLMs can inadvertently introduce factual distortions or fabricate content when processing scientific text, which could compromise dataset integrity. Our rule-based approach ensures reproducibility and maintains the original scientific meaning of the extracted evidence-conclusion pairs. After all, part of the purpose this paper is to evaluate LLMs’s capacity to digest scientific text. Below is an example evidence excerpt extracted from this process (from WGI 3.3.3.4 *Sudden Stratospheric Warming Activity*, excerpt in support of the conclusion shown in Figure 1):

Sudden stratospheric warmings (SSWs) are stratospheric weather events associated with anomalously high temperatures at high latitudes persisting from days to weeks .....

Seviour et al. (2016) found that stratosphere-resolving CMIP5 models, on average, reproduce the observed frequency of vortex splits (one form of SSWs) but with a wide range of model-specific biases .....

Some studies find an increase in the frequency of SSWs under increasing greenhouse gases .....

**Dataset Characteristics** Our extraction process yielded a compact dataset of 612 distinct “arguments” (evidence-conclusion pairs) from the IPCC AR6 reports. Each data point in our dataset contains the following features: (1) source information (Working Group report identifier and subsection header), (2) evidence bases (the supporting scientific content preceding the conclusion), (3) full conclusion paragraph, (4) individual conclusion statements (when a conclusion paragraph contains multiple assessments), and (5) the confidence level explicitly assigned to each individual conclusion statement (ranging from “low” to “high”). For the confidence classification task, we additionally created a field called “masked conclusion” where the original confidence level expressions were replaced with <MASKED>, allowing for evaluation of models’

ability to assign appropriate confidence levels without worrying about the potential bias paraphrasing introduces.

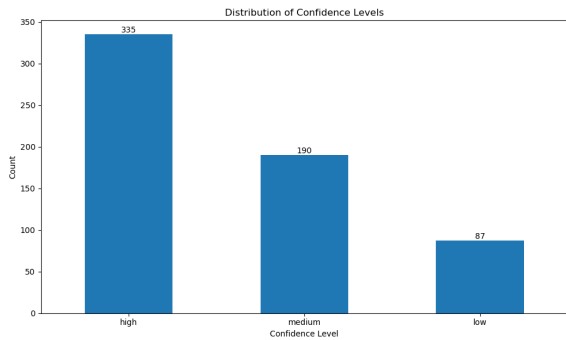


Figure 2: Confidence Level Distribution

Figure 2 shows that most conclusions have confidence, reflecting the scientific rigor of IPCC reports and the growing consensus in climate science (Cook et al., 2016). The distribution of confidence levels in our dataset is in line with what Lacombe et al. (2023); Li et al. (2024b) have observed in their climate statements datasets.

Figure 3 plots the distribution of the length of evidence texts, measured in tokens using the c110k base tokenizer, where the average length is 1654 tokens. In contrast, the average length of individual conclusion statements is only 62 tokens. This substantial difference (approximately 27:1 ratio) highlights the condensation of information required when synthesizing evidence into concise conclusions, making this a challenging task for LLMs.

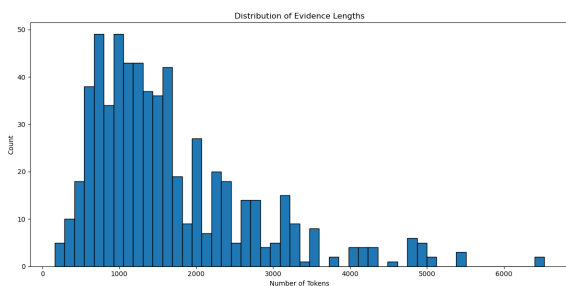


Figure 3: Evidence Length Distribution

## 4 Methods

**Contextual confidence classification** To rigorously evaluate the performance differences between reasoning-optimized LLMs and general-purpose LLMs, while also controlling for potential memorization effects, we developed three distinct prompting strategies:

1. **Zeroshot Contextual:** Models are provided with evidence bases and conclusion statements (with confidence levels masked), then asked to classify the appropriate confidence level according to IPCC standards without any examples.
2. **Fewshot Contextual:** Similar to the zero-shot approach, but with three randomly selected examples demonstrating low, medium, and high confidence classifications to provide models with context on the task.
3. **Reference Only:** Models are given only the conclusion statements, source metadata (i.e., the working group report and subsection), and standard definitions of the confidence levels—without any supporting evidence or examples. This setup serves as a control condition to test whether models are relying on memorization of the IPCC reports rather than reasoning over evidence.<sup>5</sup>

For all prompting strategies, we instructed models to select from three confidence levels ("low," "medium," or "high") based on the IPCC’s standardized confidence assessment framework (Masrandrea et al., 2011). Details about prompts are found in Appendix A.

**Factual Summarization** In the summarization task, models were given evidence bases and one example evidence-conclusion pair and then asked to generate concise summary statements that faithfully reflect the evidence while assigning appropriate confidence levels. Summaries are compared against the full conclusion, not the individual conclusions. This task evaluates models’ ability to both synthesize complex scientific information and accurately represent uncertainty—two critical components of scientific communication.

**Evaluation Metrics** For the confidence classification task, we used **accuracy** and **macro-averaged F1** score as our primary metrics. Macro-F1 is the primary metric to look at since confidence levels are somewhat imbalanced in our dataset (as shown in Figure 2).

<sup>5</sup>In the absence of a custom-trained LLM explicitly excluding IPCC AR6 materials, we concede that we cannot definitively rule out memorization. Our evaluation design instead aims to approximate this distinction by comparing performance across content-based and reference-only conditions

For factual summarization task, we adopt three commonly used metrics: <sup>6</sup>

1. **ROUGE** (Lin, 2004). ROUGE computes the overlap of n-grams between model-generated summaries and expert-written conclusions from the IPCC, providing a basic measure of content coverage. We report ROUGE-1 (unigram overlap) and ROUGE-L (longest common subsequence), using the F1 variant, which is the harmonic mean of precision (how much of the candidate matches the reference) and recall (how much of the reference is covered by the candidate).
2. **BERTScore** (Zhang et al., 2020). BERTScore improves upon ROUGE by measuring semantic similarity between generated and expert-written conclusions beyond exact word matches, using contextual embeddings from pretrained language models. We use the version based on RoBERTa-Large (Liu et al., 2019) and report the F1 score, which is standard practice in BERTScore evaluations.
3. **G-Eval** (Liu et al., 2023). with GPT-4o. G-Eval leverages LLMs with structured prompts and promises to provide human-like assessment of summary quality. We use a customized prompt tailored to our scientific evidence synthesis context to focus on relevance, faithfulness, and appropriateness of confidence levels of LLM-generated conclusions.

Unlike the evaluation of classification tasks, which benefits from clear-cut ground truth, reliable evaluation of summarization task remains an ongoing area of research (Zhang et al., 2025). We choose our evaluation metrics to balance surface-level coverage (ROUGE), semantic similarity (BERTScore), and more human-aligned quality judgments (G-Eval), given the lack of climate-specific summarization evaluation metrics. While it would be valuable

<sup>6</sup>We included FACTCC (Kryscinski et al., 2020) in earlier versions but removed it in the final version for two reasons. First, FACTCC was trained on news-style summarization datasets and may not generalize well to scientific domains like climate synthesis, where factual consistency involves nuanced reasoning and domain-specific terminology. Second, we observed potential implementation issues where FACTCC returned nearly identical scores across model outputs (up to the 4th decimal point), whereas other evaluation metrics, though close, showed more meaningful variance. This suggests that FACTCC was not a reliable discriminator in our setting.

to adapt existing metrics, such as BERTScore or FACTCC, using domain-specific models like ClimateBERT (Webersinke et al., 2022), we leave this to future work.

## 5 Classification Results

Table 1 presents the performance of both reasoning-optimized LLMs (DeepSeek-R1 and o3-mini) and a general-purpose LLM (GPT-4o) on the confidence classification task across different prompting strategies. For context, random guessing on this three-class problem would yield an expected accuracy of 33.3%, while majority class guessing (predicting "high" confidence for all examples, which constitutes approximately 55% of our dataset) would result in an accuracy of 55% with a macro-F1 score of 0.24.

**Reasoning LLMs Outperform General-Purpose Models** Both reasoning-optimized LLMs consistently outperform GPT-4o across all prompting strategies. In the zero-shot contextual setting, DeepSeek-R1 and o3-mini achieve macro-F1 scores of 65% and 63% respectively, compared to 57% for GPT-4o, representing a performance gap of 8 percentage points between DeepSeek-R1 and GPT-4o. This advantage persists in the few-shot contextual setting, where reasoning models maintain a 7 percentage point lead. The accuracy scores follow a similar pattern.

Interestingly, the few-shot approach did not consistently improve performance over zero-shot for any of the models. While o3-mini increased its F1 score from 63% to 68%, DeepSeek-R1 decreased from 65% to 63%. One potential explanation is context length limitations. Including three complete evidence-conclusion pairs in addition to the task instructions may have caused information overload, making it difficult for the models to effectively process the lengthy context.

**Memorization Is Not the Primary Driver of Performance** Given that the IPCC AR6 was published in 2023, and the knowledge cutoff dates for all tested models extend beyond this date, a natural concern is whether models are simply retrieving memorized content rather than performing genuine reasoning. The reference-only condition allows us to investigate this possibility by providing models with only the conclusion statement and retrieval-relevant information (working group and section reference) without the actual text of supporting evi-

dence.

The results reveal several important patterns. First, all models experience a performance drop in the reference-only condition compared to the contextual conditions, with GPT-4o showing the steepest decline (17 percentage points from zero-shot to reference-only). This suggests that access to evidence is indeed crucial for the task for all models. Second, even in the reference-only condition, reasoning models maintain accuracies of 57-58%, substantially above both random and majority-class baselines, while GPT-4o’s performance drops to 41%, only marginally better than a random classifier and below the majority class baseline.

The relatively strong performance of reasoning models even without evidence suggests they may be better at leveraging minimal contextual cues to retrieve information or perhaps applying general reasoning principles to scientific uncertainty assessment. However, the significant performance gap between contextual and reference-only conditions across all models indicates that genuine evidence evaluation, rather than pure memorization, drives the superior performance observed in the contextual settings.

**Performance Inference Cost Trade-off** While reasoning LLMs demonstrate superior performance, this advantage comes with significant computational costs. DeepSeek-R1 and o3-mini consume substantially more tokens during inference compared to GPT-4o, as shown in Figure 4. This difference stems from reasoning models’ explicit chain-of-thought inference-time scaling processes, where they generate extensive internal reasoning before producing a final answer.<sup>7</sup> In contrast, GPT-4o produces just 2 tokens: the prediction

<sup>7</sup>Interestingly, performance appears to correlate with tokens consumed during inference. In fewshot settings, models actually spend fewer tokens on reasoning, as if the additional input tokens from demonstrations crowded out the model’s chain-of-thoughts.

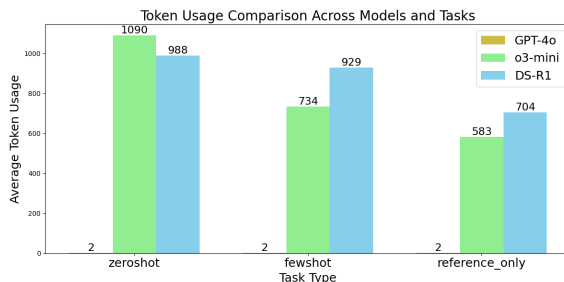


Figure 4: Token Cost Comparison

Note: The bars for GPT-4o are barely visible as it uses only 2 tokens per classification.

token and the EOS token. In practice, however, the more pressing concern is latency. Inference on DeepSeek-R1 took significantly longer than any other model, requiring over 12 hours to complete 612 requests sent asynchronously. While this largely reflects DeepSeek server’s capacity and load constraints, the pattern holds even among OpenAI models. o3-mini required approximately four times longer to complete identical tasks compared to GPT-4o.

## 6 Summarization Results

**DeepSeek-R1 seems to have a slight edge but reasoning LLMs in general do not.** As shown in Table 2, DeepSeek-R1 slightly outperforms other models on lexical and semantic similarity metrics, achieving higher scores on ROUGE-1 (0.22), ROUGE-L (0.19), and BERTScore (0.84) compared to o3-mini and GPT4o. Similarly, the differences in G-Eval are minimal. Notably, the other reasoning LLM o3-mini, while clearly outperforming GPT-4o in classification tasks, shows negligible differences in summarization performance. We are inclined to believe that reasoning LLMs may not hold a general advantage in summarization tasks, and DeepSeek-R1’s better performance may be idiosyncratic. One possible explanation for this phenomenon is that reasoning LLMs are primarily

Model	Zeroshot Contextual		Fewshot Contextual		Reference Only	
	ACC	F1	ACC	F1	ACC	F1
DS-R1	0.66	0.65	0.65	0.63	0.57	0.54
o3-mini	0.65	0.63	0.63	0.68	0.58	0.60
GPT-4o	0.58	0.57	0.57	0.56	0.41	0.41

Table 1: Classification Results. The table shows accuracy (ACC) and macro-averaged F1 (F1) scores for DeepSeek-R1, o3-mini, and GPT-4o in Zeroshot Contextual, Fewshot Contextual, and Reference only prompting settings.

Model	ROGUE-1	ROGUE-L	BERTScore	G-Eval	G-Eval	G-Eval
F1	F1	F1	F1	Faithfulness	Relevance	Confidence
DS-R1	0.22	0.19	0.84	4.80	4.90	4.94
o3-mini	0.14	0.12	0.82	4.74	4.86	4.98
GPT-4o	0.14	0.13	0.82	4.76	4.88	4.87

Table 2: Performance comparison of models on climate evidence summarization tasks. ROGUE-1 and ROUGE-L measures lexical overlap, BERTScore captures semantic similarity, and G-Eval metrics assess human-aligned quality dimensions including faithfulness, relevance, and appropriateness of confidence assessment. Higher scores indicate better performance across all metrics. Detailed evaluation prompts are provided in Appendix A.

trained to solve mathematical and logical tasks, not for open-ended, generative tasks like summarization.

**Evaluation Biases** Another possible explanation for our results lies in evaluation biases. Unlike classification tasks where evaluation is straightforward, in summarization tasks, apart from using ROGUE, we rely on pretrained language models themselves as evaluators. Recent studies such as Li et al. (2024a) and Gu et al. (2025) highlight several concerns with the use of LLMs as judges, including various forms of bias. For example, BERTScore is implemented with general-purpose pretrained language models, which are likely affected by domain shift in our climate science setting. Similarly, recent work (Panickssery et al., 2024) suggests that LLM-based evaluators may favor outputs generated by architectures similar to their own. This could partly explain why more advanced reasoning LLMs do not show clear advantages under G-Eval, especially since the evaluator used is GPT-4o itself.

That said, it is noteworthy that DeepSeek-R1, despite likely having less architectural similarity to GPT-4o than o3-mini, achieves the best overall G-Eval scores. While this complicates the interpretation, it also suggests that other factors, such as training data or output style, may influence evaluation outcomes. Addressing all of these issues is beyond the scope of this paper, and we welcome further work to develop more robust, domain-sensitive evaluation frameworks for summarization tasks.

## 7 Conclusion

Our evaluation of reasoning-optimized LLMs for climate evidence synthesis reveals both promising capabilities and important limitations. These models demonstrate significant advantages in contextual confidence classification, outperforming general-purpose LLMs by 8 percentage points in accuracy and macro-F1 scores when assigning con-

fidence levels to climate statements. This suggests potential utility as auxiliary tools for confidence assessment in scientific workflows.

However, in factual summarization tasks, reasoning LLMs show minimal and inconsistent advantages over general-purpose models. Despite their enhanced logical capabilities, they struggle with the nuanced requirements of scientific summarization when evaluated on relevance, faithfulness, confidence level assignment, which fares much worse than expert-written summaries.

These findings indicate that current reasoning LLMs can potentially contribute to specific aspects of climate evidence synthesis while highlighting the continued necessity of human expertise for summarization tasks. Future work should focus on developing specialized models for scientific synthesis and exploring human-AI collaborative frameworks that leverage the complementary strengths of both. Ultimately, a targeted approach to integrating these technologies into scientific assessment will be essential to maintain rigor while enhancing efficiency.

## 8 Limitations

We acknowledge that our research faces several limitations.

First, our evidence base excludes visual data such as graphs, charts, and images, which often contain critical climate information in IPCC reports. This omission potentially limits the comprehensiveness of our evaluation, as multi-modal reasoning capabilities would be necessary for complete assessment of climate evidence.

Second, we rely on prompt-based approaches without domain-specific adaptation or fine-tuning. While this allows for assessment of off-the-shelf model capabilities, it likely underestimates the potential performance of models specifically adapted to climate science terminology and reasoning patterns.



Third, our evaluation metrics for summarization tasks, despite careful design, may be susceptible to "LLM-as-judge" biases. Models evaluating other models' outputs could share fundamental limitations or biases, potentially inflating quality assessments of machine-generated summaries compared to expert evaluation.

Finally, our study represents a point-in-time assessment of rapidly evolving technologies. The performance gaps and capabilities identified may change significantly as reasoning LLMs continue to develop. Future work should address these limitations through multi-modal evidence inclusion, domain adaptation techniques, and more robust human-in-the-loop evaluation frameworks.

## 9 Ethical Considerations

This research evaluates LLMs on existing IPCC assessment data without involving human subjects or generating new climate recommendations. We acknowledge that AI tools for scientific synthesis raise important considerations regarding transparency, accountability, and potential automation bias. While our work demonstrates potential utility in specific tasks, we emphasize that these technologies should supplement rather than replace expert judgment in climate assessment.

## 10 Acknowledgment

This research was supported in part by credits from OpenAI's Researcher Access Program. We also gratefully acknowledge the NYU IT High Performance Computing (HPC) team for providing computational resources, services, and technical expertise that facilitated this work. We also thank the anonymous reviewers for their helpful feedback.

## References

- Thomas Bernauer and Liam F. McGrath. 2016. [Simple reframing unlikely to boost public support for climate policy](#). *Nature Climate Change*, 6(7):680–683.
- Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauss. 2024a. [Assessing large language models on climate information](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4884–4935. PMLR.
- Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Hübscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauß. 2024b. [Assessing large language models on climate information](#).
- Fengxiang Cheng, Haoxuan Li, Fenrong Liu, Robert van Rooij, Kun Zhang, and Zhouchen Lin. 2025. [Empowering llms with logical reasoning: A comprehensive survey](#).
- John Cook, Naomi Oreskes, Peter T. Doran, William R. L. Anderegg, Bart Verheggen, Ed W. Maibach, J. Stuart Carlton, Stephan Lewandowsky, Andrew G. Skuce, Sarah A. Green, et al. 2016. [Consensus on consensus: a synthesis of consensus estimates on human-caused global warming](#). *Environmental Research Letters*, 11(4):048002.
- Théo Alves Da Costa, Timothée Bohe, Jean Lelong, Nina Achache, Gabriel Olympie, Nicolas Chesneau, and Natalia De la Calzada. 2024. Climateqa, ai-powered conversational assistant for climate change and biodiversity loss.
- Adrian J. De-Gol, Corinne Le Quéré, Andrew J. P. Smith, et al. 2023. [Broadening scientific engagement and inclusivity in IPCC reports through collaborative technology platforms](#). *npj Climate Action*, 2:49.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li,

- Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. Technical report, DeepSeek-AI.
- F. M. Delgado-Chaves, M. J. Jennings, A. Atalaia, J. Wolff, R. Horvath, Z. M. Mamdouh, J. Baumbach, and L. Baumbach. 2025. *Transforming literature screening: The emerging role of large language models in systematic reviews*. *Proceedings of the National Academy of Sciences of the United States of America*, 122(2):e2411962122.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. *A survey on llm-as-a-judge*.
- Jie Huang and Kevin Chen-Chuan Chang. 2023. *Towards reasoning in large language models: A survey*. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. *A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions*. *ACM Trans. Inf. Syst.*, 43(2).
- IPCC. 2010. *Guidance note for lead authors of the IPCC fifth assessment report on consistent treatment of uncertainties*. Technical report, Intergovernmental Panel on Climate Change. Prepared by the IPCC Cross-Working Group Meeting on Consistent Treatment of Uncertainties, Jasper Ridge, CA, USA, 6-7 July 2010.
- Elphin Joe, Sai Koneru, and Christine Kirchhoff. 2024. *Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights*. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 251–257, Bangkok, Thailand. Association for Computational Linguistics.
- Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. *Evaluating the factual consistency of abstractive text summarization*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.
- Romain Lacombe, Kerrie Wu, and Eddie Dilworth. 2023. *Climatex: Do llms accurately assess human expert confidence in climate statements?* In *NeurIPS 2023 Workshop on Tackling Climate Change with Machine Learning*.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. *Llms-as-judges: A comprehensive survey on llm-based evaluation methods*.
- Ruiqi Li, Paige Reeves, Alasdair Tran, and Lexing Xie. 2024b. *Profiling and analyzing climate change statements in IPCC reports*. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. *G-eval: NLG evaluation using gpt-4 with better human alignment*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- 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*.
- Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou, editors. 2021. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to*

- the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. In press.
- Michael D. Mastrandrea, Katharine J. Mach, Gian-Kasper Plattner, et al. 2011. *The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups*. *Climatic Change*, 108:675–691.
- Amr Mohamed, Mingmeng Geng, Michalis Vazirgianis, and Guokan Shang. 2025. *Llm as a broken telephone: Iterative generation distorts information*.
- Gaku Morio and Christopher D Manning. 2023. *An nlp benchmark dataset for assessing corporate climate policy engagement*. In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.
- Veruska Muccione, Saeid Ashraf Vaghefi, Julia Bingler, et al. 2024. *Integrating artificial intelligence with expert knowledge in global environmental assessments: opportunities, challenges and the way ahead*. *Regional Environmental Change*, 24:121.
- Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. 2023. *Arabic mini-ClimateGPT : A climate change and sustainability tailored Arabic LLM*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136, Singapore. Association for Computational Linguistics.
- OpenAI. 2025. *System card: O3-Mini*. Technical report, OpenAI.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. *Llm evaluators recognize and favor their own generations*.
- Yifan Peng, Justin F. Rousseau, Edward H. Shortliffe, and Chunhua Weng. 2023. *AI-generated text may have a role in evidence-based medicine*. *Nature Medicine*, 29(7):1593–1594.
- Nick Pidgeon and Baruch Fischhoff. 2011. *The role of social and decision sciences in communicating uncertain climate risks*. *Nature Climate Change*, 1(1):35–41.
- Hans-Otto Pörtner, Debra C. Roberts, Melinda Tignor, Elvira S. Poloczanska, Katja Mintenbeck, Andrés Alegría, Morgan Craig, Stefanie Langsdorf, Sina Lösche, Vincent Möller, Andrew Okem, and Bardhyl Rama, editors. 2022. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA.
- Tobias Schimanski, Jingwei Ni, Mathias Kraus, Elliott Ash, and Markus Leippold. 2024. *Towards faithful and robust llm specialists for evidence-based question-answering*.
- P. R. Shukla, Jim Skea, Raphael Slade, Alaa Al Khourdajie, Renée van Diemen, David McCollum, Minal Pathak, Shreya Some, Purvi Vyas, Roger Fradera, Malek Belkacemi, Amrita Hasija, Giovanna Lisboa, Suvadip Luz, and Juliette Malley, editors. 2022. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK and New York, NY, USA.
- Richard C. J. Somerville and Susan Joy Hassol. 2011. *Communicating the science of climate change*. *Physics Today*, 64(10):48–53.
- Dominik Stammbach, Jingwei Ni, Tobias Schimanski, Kalyan Dutia, Alok Singh, Julia Bingler, Christophe Christiaen, Neetu Kushwaha, Veruska Muccione, Saeid A. Vaghefi, and Markus Leippold, editors. 2024. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*. Association for Computational Linguistics, Bangkok, Thailand.
- John D. Sterman. 2011. *Communicating climate change risks in a skeptical world*. *Climatic Change*, 108:811–826.
- Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, Yue Wu, Wenhai Wang, Junsong Chen, Zhangyue Yin, Xiaozhe Ren, Jie Fu, Junxian He, Wu Yuan, Qi Liu, Xihui Liu, Yu Li, Hao Dong, Yu Cheng, Ming Zhang, Pheng Ann Heng, Jifeng Dai, Ping Luo, Jingdong Wang, Ji-Rong Wen, Xipeng Qiu, Yike Guo, Hui Xiong, Qun Liu, and Zhenguo Li. 2024. *A survey of reasoning with foundation models*.
- Huy Quoc To, Ming Liu, and Guangyan Huang. 2024. *Towards efficient large language models for scientific text: A review*.
- C. W. van Eck. 2023. *The next generation of climate scientists as science communicators*. *Public Understanding of Science*, 32(8):969–984.
- Daniel Van Veen, Cornelia Van Uden, Leah Blanke-meier, Jean-Benoit Delbrouck, Ali Aali, Christian Bluethgen, Anuj Pareek, Michela Polacin, Eduardo Pontes Reis, Alena Seehofnerová, Nikita Rohatgi, Praveer Hosamani, William Collins, Nikhil Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gavidis, John Pauly, and Akshay S. Chaudhari. 2023. *Clinical text summarization: Adapting large language models can outperform human experts*. *Research Square*, pages rs.3.rs–3483777. Preprint.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan

- Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024. [Mineru: An open-source solution for precise document content extraction.](#)
- Nicolas Webersinke, Mathias Kraus, Julia Anna Binger, and Markus Leippold. 2022. [Climatebert: A pretrained language model for climate-related text.](#)
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22*, Red Hook, NY, USA. Curran Associates Inc.
- Fengli Xu, Qian Yue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. 2025. [Towards large reasoning models: A survey of reinforced reasoning with large language models.](#)
- Haopeng Zhang, Philip S. Yu, and Jiawei Zhang. 2025. [A systematic survey of text summarization: From statistical methods to large language models.](#) *ACM Comput. Surv.* Just Accepted.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert.](#) In *International Conference on Learning Representations.*

## A Appendix: Prompt Details

Following DeepSeek’s recommendation for data cleaning and general analysis tasks,<sup>8</sup> we set temperate at 1 for DeepSeek-R1 (and also GPT-4o). For o3-mini, we set reasoning effort to be “medium”.

Reference Only Prompt is adapted from Lacombe et al. (2023). Contextual prompts (zeroshot and fewshot) and summarization prompt are produced with the help of meta-prompts by OpenAI.<sup>9</sup> G4 Eval Prompt is adapted from Liu et al. (2023) and OpenAI.<sup>10</sup>

### Reference Only Prompt

You are a knowledgeable climate science assistant trained to assess the confidence level associated with various statements about climate change. You will be presented with a statement about climate science, climate impacts, or climate change mitigation retrieved from the IPCC AR6 WGI, WGII, or WGIII assessment reports.

You will be given the subsection header from where the statement is from. Climate scientists have evaluated that statement as low confidence, medium confidence, or high confidence, based on evidence (type, amount, quantity, consistency) and agreement among their peers.

What is their confidence level? Respond *only* with one of the following words: 'low', 'medium', 'high'.

Reference: {WG}, {header}  
Statement: {conclusion\_statement}  
Output:

### Contextual Prompt

Analyze the provided scientific text *evidence excerpt* and predict the <MASKED> confidence level of the *conclusion statement* based on contextual evidence.

Avoid external knowledge. Rely only on the provided text. Base your evaluation on the type, quantity, consistency, and agreement of presented evidences.

Respond *only* with one of the following words: 'low', 'medium', 'high'.

Example One

Evidence Excerpt:  
{evidence\_excerpt}  
Conclusion Statement:  
{conclusion\_statement}  
Output:  
{true\_confidence}

Example Two

Evidence Excerpt:  
{evidence\_excerpt}  
Conclusion Statement:  
{conclusion\_statement}  
Output:  
{true\_confidence}

Example Three

Evidence Excerpt:  
{evidence\_excerpt}  
Conclusion Statement:  
{conclusion\_statement}  
Output:  
{true\_confidence}

---

Input:  
Evidence Excerpt:  
{evidence\_excerpt}  
Conclusion Statement:  
{conclusion\_statement}  
Output:

<sup>8</sup>[https://api-docs.deepseek.com/quick\\_start/parameter\\_settings](https://api-docs.deepseek.com/quick_start/parameter_settings)

<sup>9</sup><https://platform.openai.com/docs/guides/prompt-generation>

<sup>10</sup>[https://cookbook.openai.com/examples/evaluation/how\\_to\\_eval\\_abstractive\\_summarization](https://cookbook.openai.com/examples/evaluation/how_to_eval_abstractive_summarization)

### Summarization Prompt

You are a scientific analyst summarizing key findings from scientific literature. Given a passage of scientific evidence, synthesize the information concisely while preserving quantitative details, uncertainty assessments, and key conclusions.

Guidelines:

1. Focus on the core scientific claims, ensuring clarity and accuracy.
2. Include key findings with numerical data and confidence levels when appropriate.
3. Be concise, your answer should not be longer than one paragraph.
4. Avoid speculation. Use only the provided information; exclude external knowledge.
5. Use precise and neutral language.

Example Input: {evidence\_excerpt}

Example Output: {conclusion}

---

Input: {evidence\_excerpt}

Output:

### G4 Eval Prompt

Scientific Conclusion Evaluation You are an expert evaluator assessing the quality of LLM-generated scientific conclusions. Your task is to evaluate how well a model has synthesized scientific literature according to specific criteria. For each submission, you will be provided with:

1. The original scientific passage
2. The LLM-generated conclusion
3. The expected guidelines for the conclusion

Evaluation Criteria (Score each on a scale of 1-5):

{criteria}

Evaluation Process: {steps}

Now evaluate: Original Passage: {passage}

LLM-Generated Conclusion: {conclusion}

{guideline\_section}

Your evaluation must follow this exact format: Evaluation:

-Relevance: Score: X/5

-Faithfulness: Score: X/5

-Confidence Level Appropriateness: Score: X/5

### G4 Eval Prompt - Relevance

Relevance

- \* 5: Perfectly captures the core scientific findings and key quantitative details
- \* 4: Identifies most important findings but misses minor details
- \* 3: Captures some key findings but omits several important elements
- \* 2: Focuses primarily on peripheral information rather than central findings
- \* 1: Fails to identify the main scientific findings

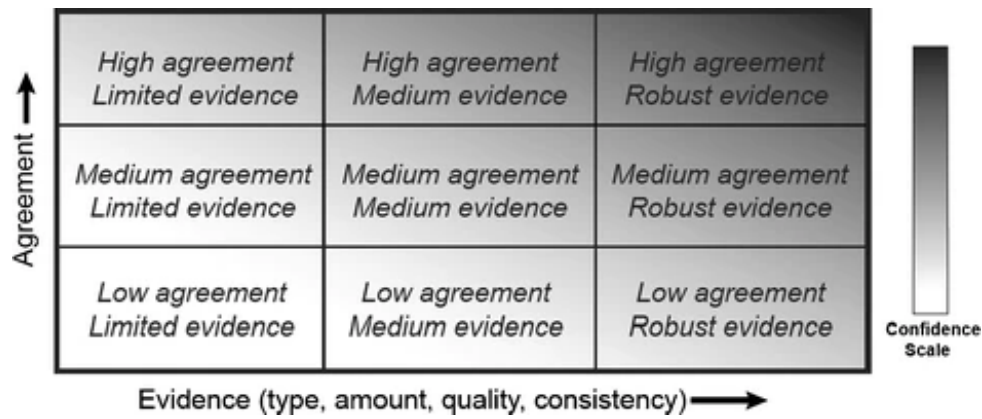


Figure 5: Confidence Evaluation Matrix from (Mastrandrea et al., 2011)

#### G4 Eval Prompt - Faithfulness

##### Faithfulness

- \* 5: Completely faithful to the original text with no misrepresentations or distortions
- \* 4: Largely faithful with only minor inaccuracies that don't affect the core meaning
- \* 3: Generally faithful but contains some misrepresentations of moderate importance
- \* 2: Contains significant misrepresentations or fabricated information
- \* 1: Fundamentally misrepresents the scientific content or contradicts the original text

#### G4 Eval Prompt - Confidence Level Appropriateness

##### Confidence Level Appropriateness

- \* 5: All confidence levels expressed in conclusion statement strictly follow from scientific text
- \* 4: Contain confidence level statements with minor inaccuracies or somewhat dubious nature
- \* 3: Preserves some uncertainty statements but omits or misrepresents others
- \* 2: Significantly understates or overstates confidence in findings
- \* 1: Completely misrepresents or omits critical uncertainty statements and confidence levels

## B Appendix: Figures from IPCC AR6

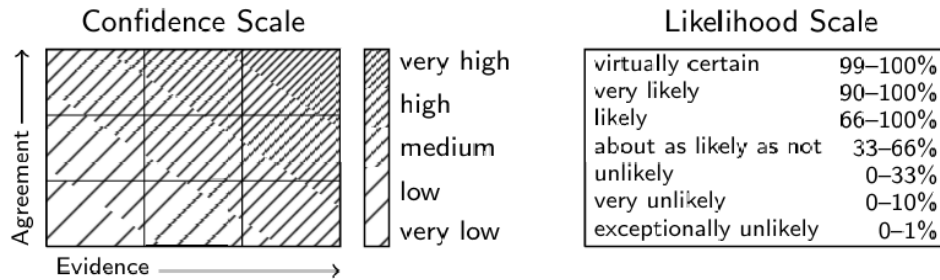


Figure 6: Confidence and likelihood scales for communicating degree of certainty in key findings of the IPCC AR5, adapted from (Mastrandrea et al., 2011)

human influence on historical blocking activity. The *low confidence* statements are due to the limited number of studies available. The shift of the Southern Hemisphere jet is correlated with modulations of the SAM (Section 3.7.2). There is *medium confidence* in model performance regarding the simulation of the extratropical jets, storm track and blocking activity, with increased resolution sometimes corresponding to better performance, but important shortcomings remain, particularly for the Euro-Atlantic sector of the Northern Hemisphere. Nonetheless, synthesizing across Sections 3.3.3.1–3.3.3.3, there is *high confidence* that CMIP6 models capture the general characteristics of the tropospheric large-scale circulation.

### 3.3.3.4 Sudden Stratospheric Warming Activity

Sudden stratospheric warmings (SSWs) are stratospheric weather events associated with anomalously high temperatures at high latitudes persisting from days to weeks. Section 2.3.1.4.5 discusses the definition and observational aspects of SSWs. SSWs are often associated with anomalous weather conditions, for example, winter cold spells, in the lower atmosphere (e.g., Butler et al., 2015; Baldwin et al., 2021).

Seviour et al. (2016) found that stratosphere-resolving CMIP5 models, on average, reproduce the observed frequency of vortex splits (one form of SSWs) but with a wide range of model-specific biases. Models that produce a better mean state of the polar vortex also tend to produce a more realistic SSW frequency (Seviour et al., 2016). The mean sea level pressure anomalies occurring in CMIP5 model simulations when an SSW is underway, however, differ substantially from those in reanalyses (Seviour et al., 2016). Unlike stratosphere-resolving models, models with limited stratospheric resolution, which make up more than half of the CMIP5 ensemble, underestimate the frequency of SSWs (Osprey et al., 2013; J. Kim et al., 2017). Taguchi (2017) found a general underestimation in CMIP5 models of the frequency of ‘major’ SSWs (which are associated with a break-up of the polar vortex), an aspect of an under-representation in those models of dynamical variability in the stratosphere. Wu and Reichler (2020) found that finer vertical resolution in the stratosphere and a model top above the stratopause tend to be associated with a more realistic SSW frequency in CMIP5 and CMIP6 models.

Some studies find an increase in the frequency of SSWs under increasing greenhouse gases (e.g., Schimanke et al., 2013; Young et al., 2013; J. Kim et al., 2017). However, this behaviour is not robust across ensembles of chemistry-climate models (Mitchell et al., 2012; Ayarzagüena et al., 2018; Rao and Garfinkel, 2021). There is an absence of studies specifically focusing on simulated trends in SSWs during recent decades, and the short record and substantial decadal variability yields *low confidence* in any observed trends in the occurrence of SSW events in the Northern Hemisphere winter (Section 2.3.1.4.5). Such an absence of a trend and large variability would also be consistent with a recent reconstruction of SSWs extending back to 1850, based on sea level pressure observations (Domeisen, 2019), although this time series has limitations as it is not based on direct observations of SSWs.

In summary, an anthropogenic influence on the frequency or other aspects of SSWs has not yet been robustly detected. There is *low confidence* in the ability of models to simulate any such trends over the historical period because of large natural interannual variability and also due to substantial common biases in the simulated mean state affecting the simulated frequency of SSWs.

## 3.4 Human Influence on the Cryosphere

### 3.4.1 Sea Ice

#### 3.4.1.1 Arctic Sea Ice

The AR5 concluded that ‘anthropogenic forcings are *very likely* to have contributed to Arctic sea ice loss since 1979’ (Bindoff et al., 2013), based on studies showing that models can reproduce the observed decline only when including anthropogenic forcings, and formal attribution studies. Since the beginning of the modern satellite era in 1979, Northern Hemisphere sea ice extent has exhibited significant declines in all months with the largest reduction in September (see Section 2.3.2.1.1, and Figures 3.20 and 3.21 for more details on observed changes). The recent Arctic sea ice loss during summer is unprecedented since 1850 (*high confidence*), but as in AR5 and SROCC there remains only *medium confidence* that the recent reduction is unique during at least the past 1000 years due to sparse observations (Sections 2.3.2.1.1 and 9.3.1.1). CMIP5 models also simulate Northern Hemisphere sea ice loss over the satellite era but with large differences among models (e.g., Massonnet et al., 2012; Stroeve et al., 2012). The envelope of simulated ice loss across model simulations encompasses the observed change, although observations fall near the low end of the CMIP5 and CMIP6 distributions of trends (Figure 3.20). CMIP6 models on average better capture the observed Arctic sea ice decline, albeit with large inter-model spread. Notz et al. (2020) found that CMIP6 models better reproduce the sensitivity of Arctic sea ice area to CO<sub>2</sub> emissions and global warming than earlier CMIP models although the models’ underestimation of this sensitivity remains. Davy and Outten (2020) also found that CMIP6 models can simulate the seasonal cycle of Arctic sea ice extent and volume better than CMIP5 models. For the assessment of physical processes associated with changes in Arctic sea ice, see Section 9.3.1.1.

Since AR5, there have been several new detection and attribution studies on Arctic sea ice. While the attribution literature has mostly used sea ice extent (SIE), it is closely proportional to sea ice area (SIA; Notz, 2014), which is assessed in Chapters 2 and 9 and shown in Figures 3.20 and 3.21. Kirchmeier-Young et al. (2017) compared the observed time series of the September SIE over the period 1979–2012 with those from different large ensemble simulations which provide a robust sampling of internal climate variability (CanESM2, CESM1, and CMIP5) using an optimal fingerprinting technique. They detected anthropogenic signals which were separable from the response to natural forcing due to solar irradiance variations and volcanic aerosol, supporting previous findings (Figure 3.21; Min et al., 2008b; Kay et al., 2011; Notz and Marotzke, 2012; Notz and Stroeve, 2016). Using selected CMIP5 models and three independently derived sets of observations, Mueller et al. (2018) detected fingerprints from greenhouse gases, natural, and other anthropogenic forcings simultaneously in the September Arctic SIE over

Figure 7: Example Section from IPCC AR6