Don't Think Twice! Over-Reasoning Impairs Confidence Calibration

Romain Lacombe [1] **Kerrie Wu** [1] **Eddie Dilworth** [1]

Abstract

Large Language Models deployed as question answering tools require robust calibration to avoid overconfidence. We systematically evaluate how reasoning capabilities and budget affect confidence assessment accuracy, using the CLIMATEX dataset (Lacombe et al., 2023a) and expanding it to human and planetary health. Our key finding challenges the "test-time scaling" paradigm: while recent reasoning LLMs achieve 48.7% accuracy in assessing expert confidence, increasing reasoning budgets consistently impairs rather than improves calibration. Extended reasoning leads to systematic overconfidence that worsens with longer thinking budgets, producing diminishing and negative returns beyond modest computational investments. Conversely, search-augmented generation dramatically outperforms pure reasoning, achieving 89.3% accuracy by retrieving relevant evidence. Our results suggest that information access, rather than reasoning depth or inference budget, may be the critical bottleneck for improved confidence calibration of knowledge-intensive tasks.

1. Introduction

The latest generation of Large Language Models (LLMs) exhibits "reasoning" abilities, a pattern of inference where models first elaborate long and intricate intermediate chains of thought, which serve as a scratchpad of sorts, before generating their final answer (Wei et al., 2023). Their widespread adoption, as tools for answering questions and orchestrating agent workflows, calls for careful evaluation of their performance under uncertainty. Calibrating the confidence of these models in particular is notoriously challenging, especially in the absence of objective ground truth as to the accuracy of statements generated in a given domain.

[1]Stanford University, Stanford, CA 94305, United States. Correspondence to: Romain Lacombe <rlacombe@stanford.edu>.

Published at ICML 2025 Workshop on Reliable and Responsible Foundation Models. Copyright 2025 by the author(s).

Accurate calibration is especially important in public-facing domains of science, from climate science to public health, where the large corpora of online text on which LLMs are trained contain long outdated and squarely incorrect content. This is particularly salient as more and more patients turn to AI systems for questions about their health, education, or other high-stakes domains.

Because climate science wrestles with daunting unknowns, from the complexity of the Earth system to the inherent uncertainty of human attempts at mitigating climate change, accurately conveying the level of confidence that experts assign to science and policy statements has long been a central task in the field (Kause et al., 2021).

This paper builds on the work by climate scientists, who meticulously labeled a vast corpus of climate-related statements with human expert confidence levels, and extends previous work by Lacombe et al. (2023a) to evaluate the calibration of the latest reasoning models to human expert confidence in statements in the climate domain.

Specifically, we rely on the CLIMATEX dataset (Expert Confidence in Climate Statements, Lacombe et al. (2023b)), a curated, expert-labeled, natural language corpus of 8,094 statements sourced from the 6th Intergovernmental Panel on Climate Change Assessment Report (IPCC AR6) (Masson-Delmotte et al., 2021; Pörtner et al., 2022; Shukla et al., 2022), and their confidence levels as assessed by scientists based on the quality and quantity of available evidence.

We use this dataset to study how recent reasoning models compare to the previously reported performance of non-reasoning LLMs on this task (Lacombe et al., 2023a). Specifically, we ask:

(i) Can LLMs accurately assess human expert confidence in climate statements? We investigate and report experimental results in Table 1.

(ii) Does test-time scaling improve confidence calibration? We evaluate models with increasing inference budgets, and report results in Figures 2 and 3.

(iii) Do our results generalize beyond climate? We introduce a novel dataset in the public health domain, and explore whether reasoning helps or impairs calibration.

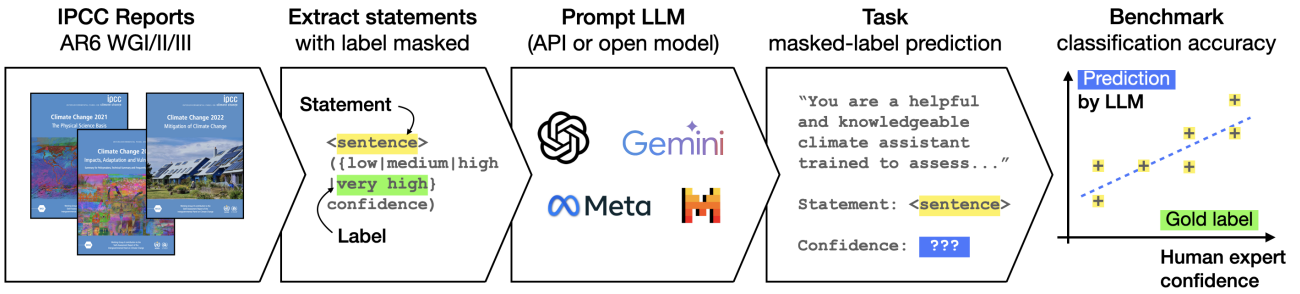

Figure 1. We use the CLIMATEX dataset (Lacombe et al., 2023b) of statements from the IPCC climate reports (Masson-Delmotte et al., 2021; Pörtner et al., 2022; Shukla et al., 2022) labeled by scientists according to a consistent, categorical confidence scale. We then evaluate and compare LLMs on their ability to reconstruct the ground truth confidence level given a statement with masked label.

2. Related work

2.1. LLMs and linguistic cues of confidence

Recent literature indicates that expressions of certainty in language influence LLM performance in calibrated NLP tasks. For instance, Zhou et al. (2023) find that appending 'weakeners' (i.e., "A: I think ...") or 'strengtheners' (i.e., "A: I'm certain...") to zero-shot question answering prompts can significantly impact LLM performance on common datasets like TriviaQA (Joshi et al., 2017). Surprisingly, strengtheners led to lower accuracy (40%) compared to weakeners (47%) in their experiments, suggesting unique difficulties in reliably interpreting linguistic cues of confidence.

In response to this issue, and to address the broader issue of LLM overconfidence, recent efforts have focused on training models to accurately convey their inherent uncertainty. Lin et al. (2022) use supervised fine-tuning (Howard & Ruder, 2018) of a large language model to elicit self-expression of how confidently it answers different arithmetic tasks, using both categorical certainty (e.g., 'high confidence') and numeric certainty (e.g., '90%'). Despite achieving promising outcomes, the authors highlighted potential issues like overfitting to the training data. Calibrating LLMs to linguistic cues of confidence in scientific domains remains a difficult open problem.

2.2. Assessing what LLMs know

Kadavath et al. (2022) demonstrated LLMs self-evaluation, showing that models can assess the veracity of their own responses against established human knowledge through few-shot learning. Their work also involved supervised fine-tuning of models to predict the likelihood of accurately answering a given question. Training and evaluation across benchmarks such as TRIVIAQA, arithmetic, and code generation revealed that while initial self-evaluation performance was low, it significantly improved with few-shot learning, using up to 20 demonstrations.

Previous work has also explored how to evaluate what LLMs "know." Chang et al. (2023) conducted a 'data archaeology' study to deduce which books LLMs were trained on, using a 'name cloze membership inference query' tasks where models are prompted to predict a masked name within a sentence based on its context. Notably, names and sentences are chosen so that human performance on this task is 0%. The authors identified a strong correlation between the frequency of books in LLM training datasets and the models' performance on the corresponding cloze task.

These results suggest that tasking LLMs to assess human expert confidence by predicting masked labels could be a challenging task, especially in the science domain, for advanced topics uncommon in LLM training sets.

2.3. LLM evaluation in the climate domain

Lastly, previous work has assessed how LLMs communicate science information, especially in the climate domain. Bulian et al. (2023) introduced a framework for evaluating the fidelity of climate-related information conveyed by LLMs. Their method involved prompting LLMs to generate extensive responses to climate questions, which were then rated by humans across dimensions including accuracy, specificity, completeness, and crucially, the appropriate communication of uncertainty levels.

IPCC reports have specifically proven to be an excellent context source for LLM systems augmented with retrieval. For instance, ChatClimate (Vaghefi et al., 2023) demonstrated the effectiveness of using scientific content to inform conversational AI agents. Earlier studies have also leveraged climate science literature as a benchmark for NLP systems, such as CLIMABENCH (Laud et al., 2023) and CLIMATEXT (Varini et al., 2021), which evaluate LLMs on tasks ranging from topic classifications to climate-related QA. Datasets such as CLIMATE-FEVER (Diggelmann et al., 2021) provide a benchmark to evaluate LLMs on the complex task of verifying climate-related assertions.

2.4. Reasoning models and test-time scaling

Much of the recent progress in LLMs has focused on "test-time scaling," the idea that allocating higher compute budgets to generating longer or searching through intermediate chains-of-thought generated by reasoning models could improve their performance (Wei et al., 2023). These techniques have demonstrated vastly improving reasoning abilities, especially in the scientific domain (Cui et al., 2025), with frontier systems such as OpenAI's `o3`, Google's `Gemini-2.5-Pro`, and Anthropic's `Claude 4` achieving accuracies comparable to human experts (OpenAI, 2025; Google, 2025; Anthropic, 2025).

Models trained to reason demonstrate improved performances in complex domains, especially in sciences (DeepSeek-AI, 2025). Reinforcement Learning with Verifiable Reward frameworks (RLVR, Lambert et al. (2025)), or Reinforcement Fine-Tuning (RFT, Luong et al. (2024)), extends these capabilities by applying policy optimization to LLM roll-outs. Supervised fine-tuning on curated reasoning traces of large models, such as the S1K dataset, also yields reasoning improvements, with scaling laws tying increased reasoning budgets to higher performance on science benchmarks (Muennighoff et al., 2025).

Test-time scaling is now broadly considered as the next paradigm to improve LLMs, and increased compute is generally expected to translate to direct, measurable gains in task performance. But does increasing the reasoning budget also lead to better calibrated models, with more accurate confidence assessments? We set out to find out.

3. Dataset

3.1. CLIMATEX: expert confidence in climate statements dataset

3.1.1. HUMAN EXPERT CONFIDENCE IN IPCC ASSESSMENT REPORTS STATEMENTS

In 2010, the IPCC issued a set of guidelines (Mastrandrea et al., 2010) to lead authors of the IPCC Reports on how to consistently communicate uncertainty. Janzwood (Janzwood, 2020) analyzes the reports written after the guidelines were published and finds evidence of broad adoption across chapter authors and IPCC reports of the 'Confidence' framework, which evaluates scientific confidence in each statement by the quality and quantity of available evidence and agreement among peers.

Confidence is measured on a 5-level categorical scale including 'very low,' 'low,' 'medium,' 'high,' and 'very high confidence' (with 'very low confidence' statements largely excluded from final reports due to lack of evidence). Adoption of this confidence framework by the large scientific community contributing to the IPCC reports makes these Assessment Reports a uniquely rich, deep, and broad corpus of statements in natural language, coherently labeled for confidence, measured as objectively as possible to reflect available evidence and scientific agreement among peers.

3.1.2. CLIMATEX DATASET CONSTRUCTIONS

This led Lacombe et al. (2023a) to introduce the CLIMATEX corpus—the Expert Confidence in Climate Statements dataset—(Lacombe et al., 2023b), comprising of 8,094 expert-annotated sentences drawn from the three most recent IPCC Assessment Report 6 volumes (Working Groups I, II, and III) (Masson-Delmotte et al., 2021; Pörtner et al., 2022; Shukla et al., 2022), to evaluate how accurately LLMs assess the confidence level attributed to climate statements by a consensus of human experts.

The dataset was constructed by extracting the complete raw text from each of the three AR6 IPCC report PDFs using an open-source library (Fenniak et al., 2022), normalizing the whitespace, tokenizing the text into sentences using NLTK (Bird et al., 2009), and using regular expression search to filter for complete sentences including a parenthetical confidence label at the end of the statement, of the form: "`<statement> ({low|medium|high|very high} confidence)`".

The complete CLIMATEX dataset contains 8,094 labeled statements, each of them labeled with the corresponding confidence rating on the IPCC's five-level scale, reflecting the evidence base and the degree of expert agreement. From these sentences, 300 randomly selected statements form a representative test dataset, while the remaining 7,794 statements form the train split.

3.2. IARC carcinogenicity expert-labeled dataset

To expand our analysis outside of the climate domain, we sought another dataset labeled for confidence by a consensus of human experts. The International Agency for Research on Cancer (IARC) reviews and evaluates the evidence on the carcinogenicity of a wide range of human exposures, for purposes of public health and cancer prevention. Based on the strength of available evidence, IARC scientist experts classify agents into one of five levels of confidence regarding their carcinogenic hazard to humans:

- Group 1: Carcinogenic to humans.
- Group 2A: Probably carcinogenic.
- Group 2B: Possibly carcinogenic.
- Group 3: Not classifiable.
- Group 4: Probably not carcinogenic.

We extracted the dataset of 1,053 agents and exposure types classified for carcinogenicity confidence in the IARC Monographs (IARC, 2025) since 1971, and used it as the basis for a masked-label classification task benchmark, to evaluate LLMs confidence calibration in the health domain.

The CLIMATEX and IARC datasets are available on HuggingFace, and the source code for experiments on GitHub.

4. Experiments

4.1. Human expert confidence prediction task

We compare the performance of recent reasoning and non-reasoning LLMs on the masked label prediction task over the CLIMATEX dataset, and evaluate them as classifiers against the ground truth labeled by human experts from IPCC reports (see results in Table 1).

Can LLMs accurately assess human expert confidence in climate statements? Using the Demonstrate-Search-Predict (DSPy) framework (Khattab et al., 2023), we presented open source LLMs and publicly available APIs for the main commercial models with sentences from the test split of the CLIMATEX dataset, and instructed them to predict the masked human expert confidence labels. We present our experimental workflow in Figure 1, and report experimental results in Table 1.

Does test-time scaling improve confidence calibration? We evaluate models on our benchmark task with increasing inference budgets, and report results in Figures 2 and 3.

Do our results generalize beyond the climate domain? We also perform the confidence label reconstruction task on the IARC carcinogenicity dataset, and report results in Figures 2 and 3.

4.2. Metrics: quantifying confidence levels

To evaluate model performance, we cast the task as four-way classification over the IPCC confidence categories ("low," "medium," "high," "very high"). We report the following complementary metrics:

- **Accuracy** — the proportion of predictions that exactly match the human-assigned label;
- **Cohen's** κ (Cohen, 1960) — a chance-corrected measurement of inter-rater agreement (here between ground truth and predicted labels), defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

  where p_o is the observed agreement and p_e is the agreement expected under random guessing across the classes (25% for our four confidence classes).
- **Average confidence** — the average predicted score on a dimensionless confidence scale (0.0: no confidence – 3.0: highest confidence).

Table 1 presents accuracy and Cohen's κ for every experiment with reasoning and non-reasoning models.

We report 95% confidence intervals for accuracy and confidence, using bootstrapping with resampling of our experimental results.

4.3. Baselines

4.3.1. NON-EXPERT HUMAN BASELINE

To establish a non-expert human baseline, Lacombe et al. (2023a) report an experiment with 3 college-educated, non-climate expert participants, who were presented with the statements from the CLIMATEX test dataset and asked to classify them by their best guess of the consensus confidence level (low, medium, high, or very high). They achieved an accuracy of 36.2% and Cohen's κ of 14.9% in aligning their estimated confidence assessments with the ground truth provided by climate science experts. This baseline serves as a loose reference point to gauge the performance of LLMs, and highlights the complexity of the task for humans.

4.3.2. CLASSIFIER BASELINE

To establish a baseline for the accuracy achievable through full gradient-descent training, we developed a classifier baseline using the RoBERTa architecture (Liu et al., 2019b). We fine-tuned the pre-trained `RoBERTa-Large` model on the CLIMATEX train set for 2 epochs, with over-sampling of low and very high confidence statements to match the test set distribution. The fine-tuned RoBERTa classifier achieved an accuracy of 53.7% and Cohen's κ of 38.3% for prediction of the expert-assigned confidence levels on the test set.

4.4. Experimental controls

4.4.1. EXTERNAL VALIDITY

To control for the possibility that LLMs are following our confidence assessment instructions randomly, we conducted experiments using a dataset of statements from The Onion (Thomas, 2023), a satirical news website, and grammatically correct but non-sensical sentences from the ReCOGS NLP dataset (Wu et al., 2023). Models declined to provide an answer at relatively high rates by responding "I don't know," suggesting that labeling results on CLIMATEX are not simply random predictions.

We also verified external validity of the model prompt on statements outside of the CLIMATEX dataset. We ran the masked confidence label prediction task on the FEVER (Thorne et al., 2018) True and False sentences, and obtain

Model	Accuracy	Cohen's κ	Bias	Parameters
Search-Augmented Models				
Google Gemini 2.5 Pro with Search	**89.3%**	**85.7%**	+0.030	Unknown
Google Gemini 2.5 Flash with Search	88.3%	84.4%	+0.097	Unknown
Reasoning Models				
Google Gemini 2.5 Pro	**48.7%**	**31.6%**	+0.066	Unknown
Google Gemini 2.5 Pro – Bulk processing	45.3%	27.1%	+0.353	Unknown
Google Gemini 2.5 Flash – Best thinking budget	45.0%	26.7%	+0.265	Unknown
OpenAI o3 – Program synthesis	40.7%	20.9%	+0.167	Unknown
Non-Reasoning Models				
Google Gemini 1.5 Pro	**45.0%**	**26.7%**	+0.230	Unknown
OpenAI GPT-4o	44.0%	25.3%	+0.283	Unknown
OpenAI GPT-4	42.4%	23.2%	+0.197	Unknown
OpenAI GPT-3.5 Turbo	39.7%	19.6%	+0.226	Unknown
Open-Source LLMs				
Meta Llama 3 8B Chat	**41.1%**	**21.5%**	-0.001	8B
Mixtral-8x22B Instruct v0.1	38.1%	17.1%	+0.418	8×22B
Meta Llama 3 70B Chat	36.2%	14.9%	+0.444	70B
Mixtral-8x7B Instruct v0.1	35.9%	14.5%	+0.303	8×7B
Mistral 7B Instruct v0.3	35.0%	13.3%	+0.423	7B
Google Gemma Instruct 2B	33.9%	11.9%	+0.010	2B
Google Gemma Instruct 7B	33.4%	11.2%	+0.305	7B
Baselines				
RoBERTa-Large fine-tuned	**53.7%**	**38.3%**	–	355M
Non-expert humans	36.2%	14.9%	–	60T

Table 1. Experimental results: accuracy and Cohen's κ for the masked confidence label prediction task on CLIMATEX.

higher average confidence classification on true statements than on false statements, which serves as a robustness check.

Lastly, the higher confidence obtained on the IARC carcinogenicity dataset (60-66%, cf. Figure 3) suggests that our results generalize beyond the climate domain.

4.4.2. SELF-CONSISTENCY AND ROBUSTNESS TO TEMPERATURE

To evaluate the self-consistency of LLM answers and their robustness to sampling randomness, we conducted multiple runs of the confidence assessment task on the CLIMATEX dataset using a temperature of 1. Results for each sentence were largely consistent across runs. For maximum reproducibility, we ran subsequent experiments with a temperature of 0 whenever possible.

4.4.3. ROBUSTNESS TO PROMPT WORDING

To investigate whether LLMs react to confidence levels implicitly expressed in the prompt template, we varied the prompts used to elicit confidence predictions from the LLMs, while keeping the underlying CLIMATEX dataset unchanged. Results were largely unchanged, which confirms that differences between models are a reflection of their capability rather than their prompts.

5. Results and analysis

5.1. Confidence calibration

5.1.1. RECENT LLMS CAN CLASSIFY HUMAN EXPERT CONFIDENCE IN CLIMATE-RELATED STATEMENTS

Our experiments extend and update results from Lacombe et al. (2023a) and show how accurately recent LLMs can assess human expert confidence in climate statements: the best

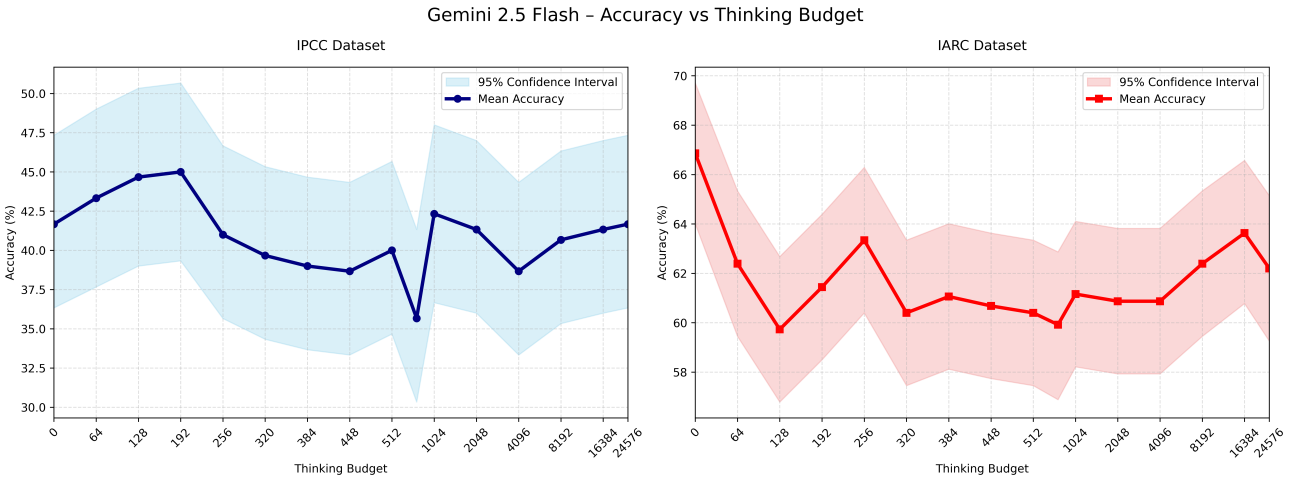

Figure 2. Accuracy (%) vs thinking budget (tokens) for Gemini 2.5 Flash.

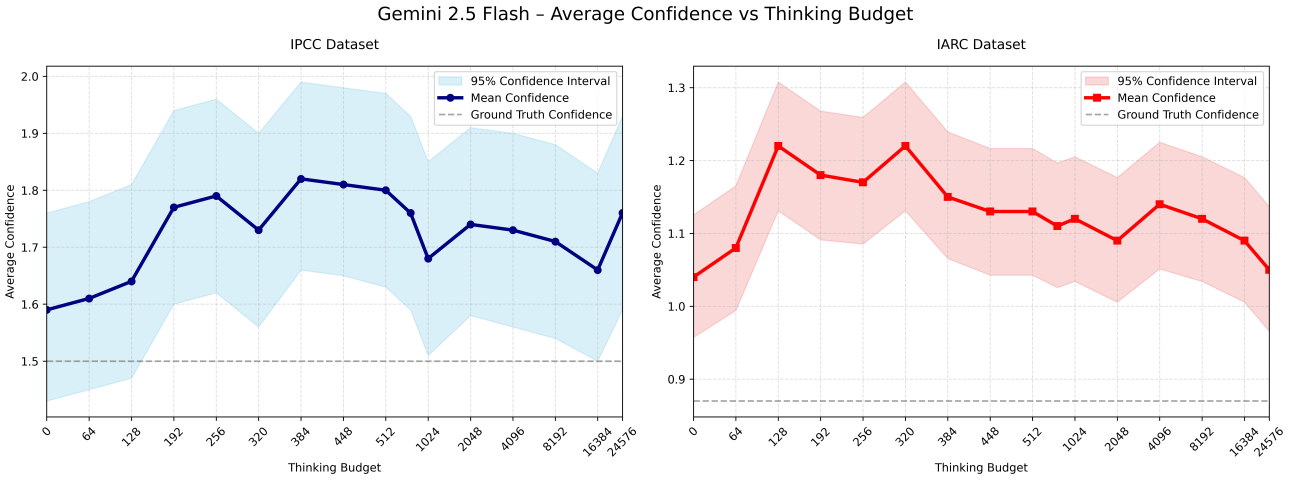

Figure 3. Confidence (dimensionless scale [0.0–3.0]) vs thinking budget (tokens) for Gemini 2.5 Flash.

performing LLM on our benchmark, `Gemini-2.5-Pro` can classify masked-label statements in the CLIMATEX test set with up to 48.7% accuracy against human expert labels, corresponding to a Cohen's κ of 31.6%.

Although the confidence prediction task remains challenging for recent LLMs, **the performance of state-of-the-art models has markedly improved** compared to early 2023 (39.7% for GPT-3.5), and is undeniably better than the 36.2% non-expert human baseline. However, it still falls short of the performance of a very small (355M params) encoder model fine-tuned on the CLIMATEX training set.

Importantly, all models exhibit a positive correlation between ground truth and predicted confidence level, with a slope varying from 0.049 for `Gemma-Instruct-7B` (near random classifier), up to 0.360 for `Mixtral-8x22B-Instruct-v0.1`, and classification accuracy that is significantly higher than random chance (25%).

5.1.2. MOST MODELS ARE CONSISTENTLY OVERCONFIDENT

For a quantitative analysis of confidence calibration, we map IPCC confidence levels to a numerical score, where 'low' corresponds to 0, 'medium' to 1, 'high' to 2, and 'very high' to 3. This mapping allows us to treat our LLM classifiers as regression models, and assess their performance by contrasting predicted confidence scores to ground truth numerical values as in a regression setting.

Meta AI's `Llama-3-70B-Chat` performs worse on this dimension, consistently making strongly overconfident classifications (+0.444 bias). In contrast with `Llama-3-8B-Chat`, which exhibits lower bias (close to

0), all other models consistently overestimate certainty level of the statements with which they are presented. Importantly, models specifically overestimate confidence in the 'low' and 'medium' categories within the CLIMATEX dataset, which are the most inherently uncertain.

5.1.3. SIZE IS NOT DESTINY

Closed-sourced LLMs are presumed to have large numbers of parameters and to have been pre-trained on vast corpora with vastly larger quantities of compute than currently available open-source models.

Unsurprisingly, the former generally perform better than the latter. The best open-source model (`Llama-3-8B-Chat`) only out-classed the worst performing closed-source LLM (`GPT-3.5-Turbo`), which suffers from the additional limitation of a knowledge cutoff date (September 2021) predating the release of the last two IPCC AR6 reports in 2022.

However, among models with known parameter counts, **model size did not always correlate with improved confidence calibration**. For instance, `Llama-3-8B-Chat` beats `LLama-3-70B-Chat`, which exhibits excessive over-confidence, by a large margin; or `Gemma-Instruct-2B` performs marginally better than the 3.5 times larger `Gemma-Instruct-7B`.

5.2. Reasoning and emerging capabilities

State-of-the-art LLMs offer several orthogonal test-time compute levers to scale performance:

- **Deliberation**: the ability to allocate more internal tokens to "reasoning" steps (e.g., `Gemini 2.5 Flash`'s 'thinking budget' feature);
- **Context length**: windows large enough to hold hundreds of statements or even entire corpora and process them at once in bulk.
- **Tool use**: the ability to execute code or call tools such as search engines, to improve completions.

We investigate how the first two levers affect confidence calibration, and explore how tool use—specifically Search for retrieval-augmented generation (RAG)—that explicitly injects external evidence in the model context.

5.2.1. REASONING AND CONFIDENCE CALIBRATION

Does "thinking longer" improve confidence accuracy? In Figure 2, we report the accuracy achieved by `Gemini 2.5 Flash` for masked label prediction on the CLIMATEX test set and the IARC dataset against reasoning budget, as we perform a sweep of the thinking budget hyperparameter from 0 to 24,576 (max) tokens.

Three trends emerge:

- **Limited early gains.** A modest budget (64–192 tokens) lifts accuracy from 41.7% (single-step inference) to a peak of 45.0% on the IPCC dataset, a +3.3 points jump that mirrors the "chain-of-thought" effect reported by researchers (Wei et al., 2023).
- **Diminishing—and eventually negative—returns.** Beyond 256 tokens, accuracy collapses, bottoming out near 35.7% at 768 tokens. The model appears to over-reason, and longer chains of thought may introduce spurious rationales or circular reasoning patterns that hurt both accuracy and calibration. This effect is immediate for the IARC dataset, where introducing reasoning drops performance from 66.9% to 62.4%.
- **Partial recovery at extreme budgets.** Pushing past a reasoning budget of 8,192 tokens slowly restores performance, but even the maximum budget ends only slightly above the baseline (41.7% at 24,576 tokens for IPCC, and 62.2% for IARC).

Does "thinking longer" improve confidence calibration? In Figure 3, we report the average confidence for masked label prediction by `Gemini 2.5 Flash` on the CLIMATEX test set and the IARC dataset against reasoning budget, as we perform a sweep of the thinking budget hyperparameter from 0 to 24,576 (max) tokens. We compute confidence according to the mapping from section 5.1.2, and report the average ground truth confidence level, which an unbiased model should achieve.

Surprisingly, models exhibit increasing and marked overconfidence as the reasoning budget grows. For instance, `Gemini 2.5 Flash` grows from +6% overconfident to +21.3% overconfident on the IPCC dataset as reasoning budget grows from 0 to 384 tokens, and from +15.6% to +35.6% overconfident on the IARC dataset when extending thinking budget to 320 tokens.

Overall, larger reasoning budgets for longer chains of thought appear to impair confidence calibration instead of improving it. These results overall point at a cautionary message for "test-time scaling": more inference-time computation alone does not automatically elicit better factuality or improve model confidence calibration to human experts.

5.2.2. LONG CONTEXT AND PROGRAM SYNTHESIS

The large context windows of recent models enable longer tasks requiring much larger token numbers. This lets us bypass statement-by-statement prompting. We experimented with long context tasks by feeding `Gemini 2.5 Pro`'s 1M-token window the entire 300-statement test split in a single prompt, asking for one-shot confidence labeling.

The model returned predictions for every line with 45.3% accuracy—exceeding the statement-by-statement score of `Gemini 1.5 Pro`—while also producing well-formatted CSV output that worked seamlessly with our evaluation script.

An even more surprising emergent behavior surfaced in `OpenAI o3`. When supplied the same CSV file, o3 refused to label all 300 statements directly, arguing that the task would be too long. When prompted repeatedly, the model eventually generated a concise Python script to run a smart labeling heuristic based on the presence of certain words (e.g., 'high' confidence for sentences including 'very likely,' 'medium' for 'suggest,' 'low' for 'unlikely').

Astonishingly, running the generated script yielded 40.7% accuracy—lower than the top performing models, but remarkable nonetheless, especially since (i) the model wrote the program unprompted, and (ii) this heuristic exceeds the performance of non-expert humans on the task.

This behavior hints at a form of latent program-synthesis fallback: when unable to perform the task directly, the model generates an algorithm to solve it programmatically instead. This emerging capability warrants future research and investigation.

5.2.3. Search-augmented generation

Finally, we explored retrieval-augmented generation by contextualizing each statement with evidence from an online search performed by the model. This simple RAG setup essentially solved the task close to perfection, and largely saturated our benchmark. With `Gemini 2.5 Pro` (`Flash` respectively), search-augmented generation improved accuracy from 48.7% to 89.3% (45.0% to 88.3%), and Cohen's κ from 31.6% to 85.7 % (26.7% to 84.4%).

These impressive gains suggest that LLM understanding of complex scientific claims is, to a surprising extent, **bottlenecked by access to the right evidence rather than by the model's innate reasoning capacity**. Once the retrieval component surfaces salient passages, the model assigns the correct categorical confidence almost every time.

These findings echo results from open-book question answering and suggests that, for knowledge-intensive tasks, further scaling of raw model size or "thinking budget" may yield diminishing returns relative to investments in retrieval depth and breadth. They also call into question what reasoning capabilities benchmarks actually measure: difficult evaluation sets which allow for Search, such as GPQA (Rein et al., 2023), typically do not control for the quality of retrieved evidence, and risk confusing reasoning deficiencies with limitations of knowledge retrieval.

5.3. Limitations and future work

This study acknowledges several limitations that could impact the generalizability of our findings.

5.3.1. Bias from LLM pre-training data

Results reported in the IPCC reports are generally well established and therefore extensively published in the literature. As a consequence, they are present in the large corpora of crawled text used for training LLMs, which raises the question of whether benchmark results merely reflect inputs to LLM pre-training, rather than model capabilities.

This is particularly relevant for fairly comparing model performance—since some models (i.e., `GPT-3.5-Turbo`) have knowledge cutoff predating the last two IPCC reports, while other models have cutoff dates posterior to all three IPCC reports. Further experiments could use data ablation studies (Liu et al., 2019a), or evaluation on held-out unpublished research (e.g., like SWE-Bench (Jimenez et al., 2024)), to disentangle these effects.

5.3.2. Free-form calibration

Calibrating models accurately to real-world knowledge and confidence levels is critical for LLMs, with increasingly high stakes as these models are deployed for scientific research, code-generation agents, or information processing workflows with potentially large reach.

Quantifying how accurately LLM generations match ground truth confidence is an important contribution to address this issue. However, our current approach relies on using the LLM as a classifier rather than in free-form generation. Further work could study how accurately encoder-based models, such as the classifier we trained as our baseline, rate the accuracy of LLMs confidence in longer outputs.

5.3.3. Evolving confidence baselines

The human expert confidence labels we use as our ground truth are inherently a moving target, formed through consensus based on available evidence at a given time. Future work could study how these confidence assessments evolve over time as new literature and evidence emerge.

The dynamic nature of confidence levels underscores the need for ongoing evaluation and adaptation of NLP models and benchmarks (Kiela et al., 2021). Notably, AI for Science would benefit from a better understanding of how to update implicit confidence levels on the basis of new experimental results.

5.3.4. Limited human baselines

The human baseline reported by Lacombe et al. (2023a) was limited to three non-expert participants, and may be biased

as it was performed while finalizing the CLIMATEX test set. Future work could include an updated baseline with more participants for better statistical representativeness, and include the IARC dataset.

The baseline also leaves out the question of how expert humans themselves perform on this task. Given the complexity and the multidisciplinary nature of climate sciences, it is possible that even domain experts would not perform well on our benchmark. Better understanding the limits of human expert performance would inform future developments of new LLM evaluations building on CLIMATEX, and future work could include a baseline study with domain experts, such as authors and chapter leads of the IPCC reports.

6. Conclusion

This study fundamentally challenges the assumption that more reasoning automatically leads to better-calibrated AI systems. Our systematic evaluation using expert-labeled climate science statements reveals that extended reasoning consistently impairs rather than improves confidence calibration, leading to systematic overconfidence that worsens with longer thinking budgets.

Calibrating the frequently "confidently wrong" LLMs is critical for language model-based applications to function effectively as knowledge retrieval systems. The ability to benchmark whether generated outputs adequately convey the underlying ground truth confidence is paramount for LLMs to meet that objective.

This is especially true for reasoning models, which are now often forming the basis for multi-agent systems and modular architectures that rely on the accuracy of the underlying model to perform complex tasks. Evaluating, measuring, and improving their confidence calibration is therefore critical for the fields of reasoning agents, large language models, and Machine Learning systems.

In this work, we've explored whether "test-time scaling" and increased inference time compute budgets help the latest generation of reasoning models improve their factuality and their confidence calibration. **The answer is "no": while our benchmark is topped by the reasoning models, increasing their thinking budget makes models overconfident and impairs their confidence calibration!**

To reach these conclusions, we expanded on previous work and used the CLIMATEX dataset, a curated, natural language corpus of 8,094 statements sourced from the 6th IPCC Assessment Report, expert-labeled for confidence levels by climate scientists. We also expended our confidence calibration benchmark to public health, epidemiology, oncology, and toxicology, with the IARC dataset of agent exposure carcinogenicity confidence.

To our knowledge, ours is the first study to systematically document calibration of LLMs to human expert confidence levels in reasoning models, and to investigate how reasoning budget impacts confidence calibration in the human and planetary health domain.

Still, a key result we report is a negative one: we show that increasing reasoning budget through "test-time scaling" impairs the confidence calibration of reasoning models.

Our finding that search-augmented generation dramatically outperforms even the best reasoning models (89.3% vs 48.7% accuracy) points toward a more promising path forward. Rather than scaling reasoning compute, practitioners should prioritize improving information retrieval and evidence synthesis capabilities. This aligns with broader trends in AI research suggesting that access to relevant information often matters more than raw computational power for knowledge-intensive tasks.

Acknowledgments

The authors wish to thank Christopher Potts and Mina Lee. We are also grateful to the thousands of scientists and experts who contributed to the IPCC AR6 reports and the IARC monographs, and the many more whose research informed them; this work would not have been possible without them.

References

Anthropic. System card: Claude Opus 4 and Sonnet 4, 2025. URL https://www.anthropic.com/news/claude-4.

Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009. URL https://www.oreilly.com/library/view/natural-language-processing/9780596516499/.

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., Ciaramita, M., Gaiarin, B., Huebscher, M. C., Buck, C., Mede, N., Leippold, M., and Strauss, N. Assessing Large Language Models on climate information, 2023. URL https://arxiv.org/abs/2310.02932.

Chang, K. K., Cramer, M., Soni, S., and Bamman, D. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4, 2023. URL https://arxiv.org/abs/2305.00118.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46, 1960. doi: 10.1177/001316446002000104. URL https://journals.sagepub.com/doi/10.1177/001316446002000104.

Cui, H., Shamsi, Z., Cheon, G., Ma, X., Li, S., Tikhanovskaya, M., Norgaard, P., Mudur, N., Plomecka, M., Raccuglia, P., Bahri, Y., Albert, V. V., Srinivasan, P., Pan, H., Faist, P., Rohr, B., Cubuk, E. D., Aykol, M., Merchant, A., Statt, M. J., Morris, D., Purves, D., Kleeman, E., Alcantara, R., Abraham, M., Mohammad, M., VanLee, E. P., Jiang, C., Dorfman, E., Kim, E.-A., Brenner, M. P., Jain, V., Ponda, S., and Venugopalan, S. CURIE: Evaluating LLMs On Multitask Scientific Long Context Understanding and Reasoning, 2025. URL https://arxiv.org/abs/2503.13517.

DeepSeek-AI, Daya Guo, D. Y. H. Z. J. S. R. Z. e. a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2501.12948.

Diggelmann, T., Boyd-Graber, J., Bulian, J., Ciaramita, M., and Leippold, M. Climate-fever: A dataset for verification of real-world climate claims, 2021. URL https://arxiv.org/abs/2012.00614.

Fenniak, M., Stamy, M., pubpub zz, Thoma, M., Peveler, M., exiledkingcc, and PyPDF2 Contributors. The PyPDF2 library, 2022. URL https://pypi.org/project/PyPDF2/.

Google. Gemini 2.5 Pro System Card, 2025. URL https://storage.googleapis.com/model-cards/documents/gemini-2.5-pro.pdf.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification, 2018. URL https://arxiv.org/abs/1801.06146.

IARC. IARC monographs on the identification of carcinogenic hazard to humans, 2025. URL https://monographs.iarc.who.int/list-of-classifications/.

Janzwood, S. Confident, likely, or both? The implementation of the uncertainty language framework in IPCC special reports. *Climatic Change*, 162(3):1655–1675, October 2020. ISSN 1573-1480. doi: 10.1007/s10584-020-02746-x. URL https://doi.org/10.1007/s10584-020-02746-x.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. SWE-bench: Can language models resolve real-world github issues? In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66. arXiv:2310.06770.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., Ganguli, D., Hernandez, D., Jacobson, J., Kernion, J., Kravec, S., Lovitt, L., Ndousse, K., Olsson, C., Ringer, S., Amodei, D., Brown, T., Clark, J., Joseph, N., Mann, B., McCandlish, S., Olah, C., and Kaplan, J. Language Models (Mostly) Know What They Know, 2022. URL https://arxiv.org/abs/2207.05221.

Kause, A., de Bruin, W. B., Domingos, S., Mittal, N., Lowe, J., and Fung, F. Communications about uncertainty in scientific climate-related findings: a qualitative systematic review. *Environmental Research Letters*, 16(5):053005, 2021. doi: 10.1088/1748-9326/abb265. URL https://iopscience.iop.org/article/10.1088/1748-9326/abb265.

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., and Zaharia, M. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP, 2023. URL https://arxiv.org/abs/2212.14024.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., Ma, Z., Thrush, T., Riedel, S., Waseem, Z., Stenetorp, P., Jia, R., Bansal, M., Potts, C., and Williams, A. Dynabench: Rethinking benchmarking in nlp, 2021. URL https://arxiv.org/abs/2104.14337.

Lacombe, R., Wu, K., and Dilworth, E. ClimateX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements? *Tackling Climate Change with Machine Learning workshop at NeurIPS 2023*, 2023a. URL https://arxiv.org/abs/2311.17107.

Lacombe, R., Wu, K., and Dilworth, E. Expert Confidence in Climate Statements (CLIMATEX) dataset, 2023b. URL https://huggingface.co/datasets/rlacombe/ClimateX.

Lambert, N., Morrison, J., Pyatkin, V., Huang, S., Ivison, H., Brahman, F., Miranda, L. J. V., Liu, A., Dziri, N., Lyu, S., Gu, Y., Malik, S., Graf, V., Hwang, J. D., Yang, J., Bras, R. L., Tafjord, O., Wilhelm, C., Soldaini, L., Smith, N. A., Wang, Y., Dasigi, P., and Hajishirzi, H. Tulu 3: Pushing Frontiers in Open Language Model Post-Training, 2025. URL https://arxiv.org/abs/2411.15124.

Laud, T., Spokoyny, D., Corringham, T., and Berg-Kirkpatrick, T. ClimaBench: A Benchmark Dataset For

Climate Change Text Understanding in English, 2023. URL https://arxiv.org/abs/2301.04253.

Lin, S., Hilton, J., and Evans, O. Teaching Models to Express Their Uncertainty in Words, 2022. URL https://arxiv.org/abs/2205.14334.

Liu, N. F., Schwartz, R., and Smith, N. A. Inoculation by fine-tuning: A method for analyzing challenge datasets, 2019a. URL https://arxiv.org/abs/1904.02668.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019b. URL https://arxiv.org/abs/1907.11692.

Luong, T. Q., Zhang, X., Jie, Z., Sun, P., Jin, X., and Li, H. ReFT: Reasoning with Reinforced Fine-Tuning, 2024. URL https://arxiv.org/abs/2401.08967.

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B. (eds.). *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, 2021. doi: 10.1017/9781009157896. URL https://www.ipcc.ch/report/ar6/wg1/.

Mastrandrea, M. D., Field, C. B., Stocker, T. F., Edenhofer, O., Ebi, K. L., Frame, D. J., Held, H., Kriegler, E., Mach, K. J., Matschoss, P. R., Plattner, G.-K., Yohe, G. W., and Zwiers, F. W. Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties, 2010. URL https://www.ipcc.ch/site/assets/uploads/2017/08/AR5_Uncertainty_Guidance_Note.pdf.

Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. s1: Simple test-time scaling, 2025. URL https://arxiv.org/abs/2501.19393.

OpenAI. OpenAI o3 and o4-mini system card, 2025. URL https://openai.com/index/o3-o4-mini-system-card/.

Pörtner, H. O., Roberts, D. C., Tignor, M., Poloczanska, E. S., Mintenbeck, K., Alegría, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., and Rama, B. (eds.). *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, USA, 2022. ISBN 9781009325844. doi: 10.1017/9781009325844. URL https://www.ipcc.ch/report/ar6/wg2/.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof qa benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Shukla, P. R., Skea, J., Slade, R., Al Khourdajie, A., van Diemen, R., McCollum, D., Pathak, M., Some, S., Vyas, P., Fradera, R., Belkacemi, M., Hasija, A., Lisboa, G., S., L., and J., M. (eds.). *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, USA, 2022. ISBN 9781009157926. doi: 10.1017/9781009157926. URL https://www.ipcc.ch/report/ar6/wg3/.

Thomas, B. Onion News Articles Dataset, 2023. URL https://huggingface.co/datasets/Biddls/Onion_News.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018. URL https://aclanthology.org/N18-1074/.

Vaghefi, S. A., Wang, Q., Muccione, V., Ni, J., Kraus, M., Bingler, J., Schimanski, T., Colesanti-Senni, C., Webersinke, N., Huggel, C., and Leippold, M. chatClimate: Grounding conversational AI in climate science, 2023. URL https://arxiv.org/abs/2304.05510.

Varini, F. S., Boyd-Graber, J., Ciaramita, M., and Leippold, M. Climatext: A dataset for climate change topic detection, 2021. URL https://arxiv.org/abs/2012.00483.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Wu, Z., Manning, C. D., and Potts, C. ReCOGS: How incidental details of a logical form overshadow an evaluation of semantic interpretation, 2023. URL https://arxiv.org/abs/2303.13716.

Zhou, K., Jurafsky, D., and Hashimoto, T. Navigating the Grey Area: Expressions of Overconfidence and Uncertainty in Language Models, 2023. URL https://arxiv.org/abs/2302.13439.