

Stanford



Limits of Reasoning Models in Science

Summer 2025 report card! :)

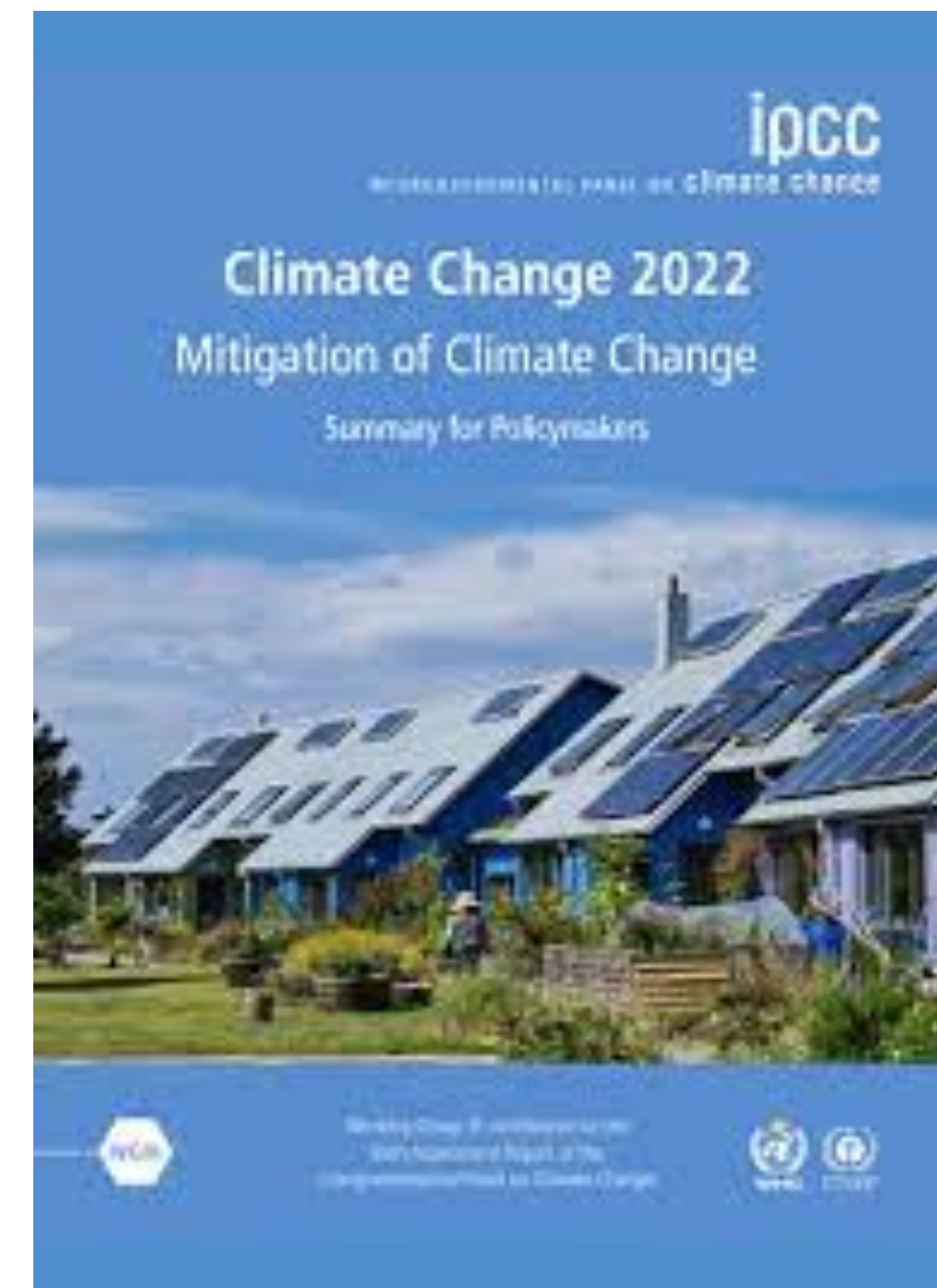
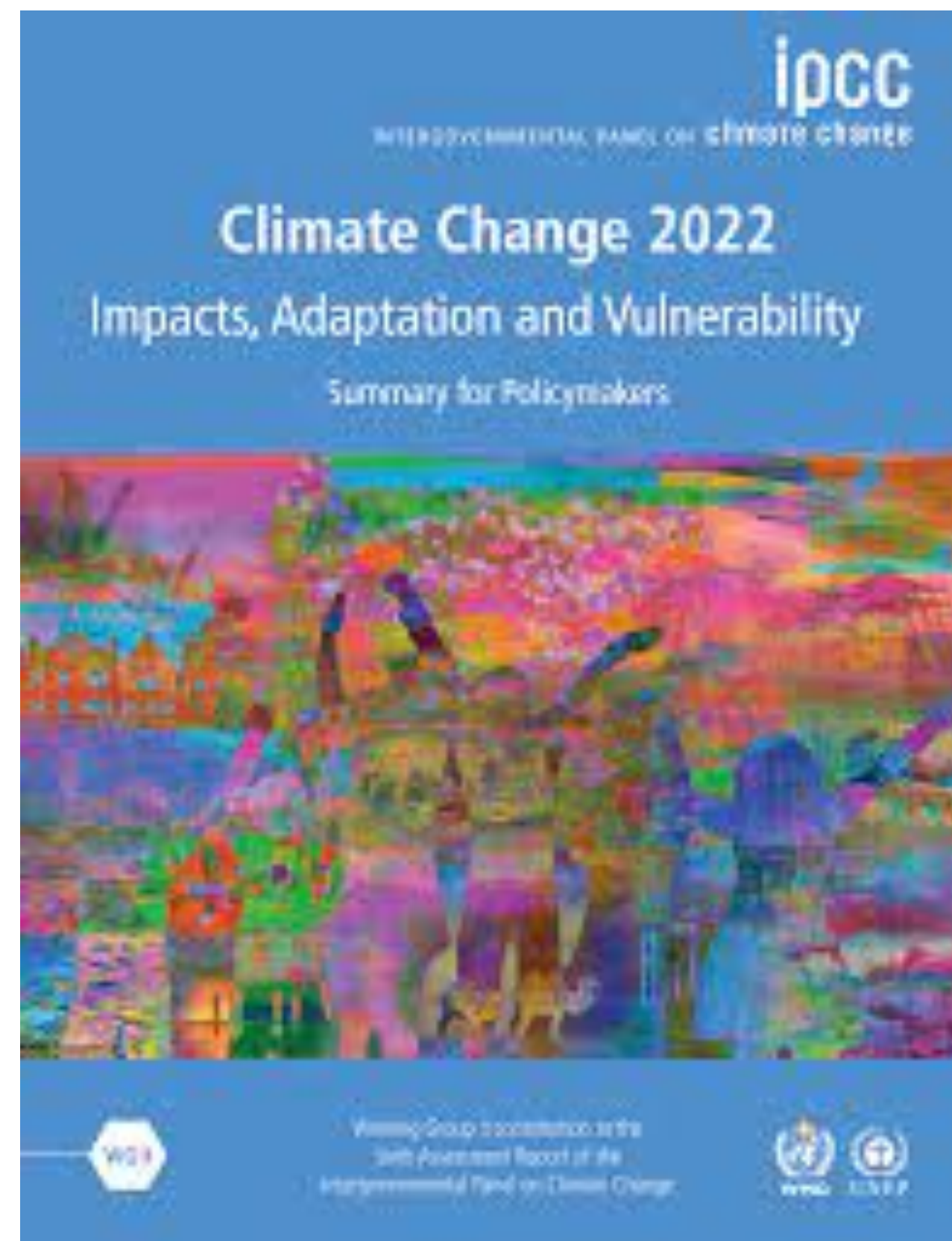
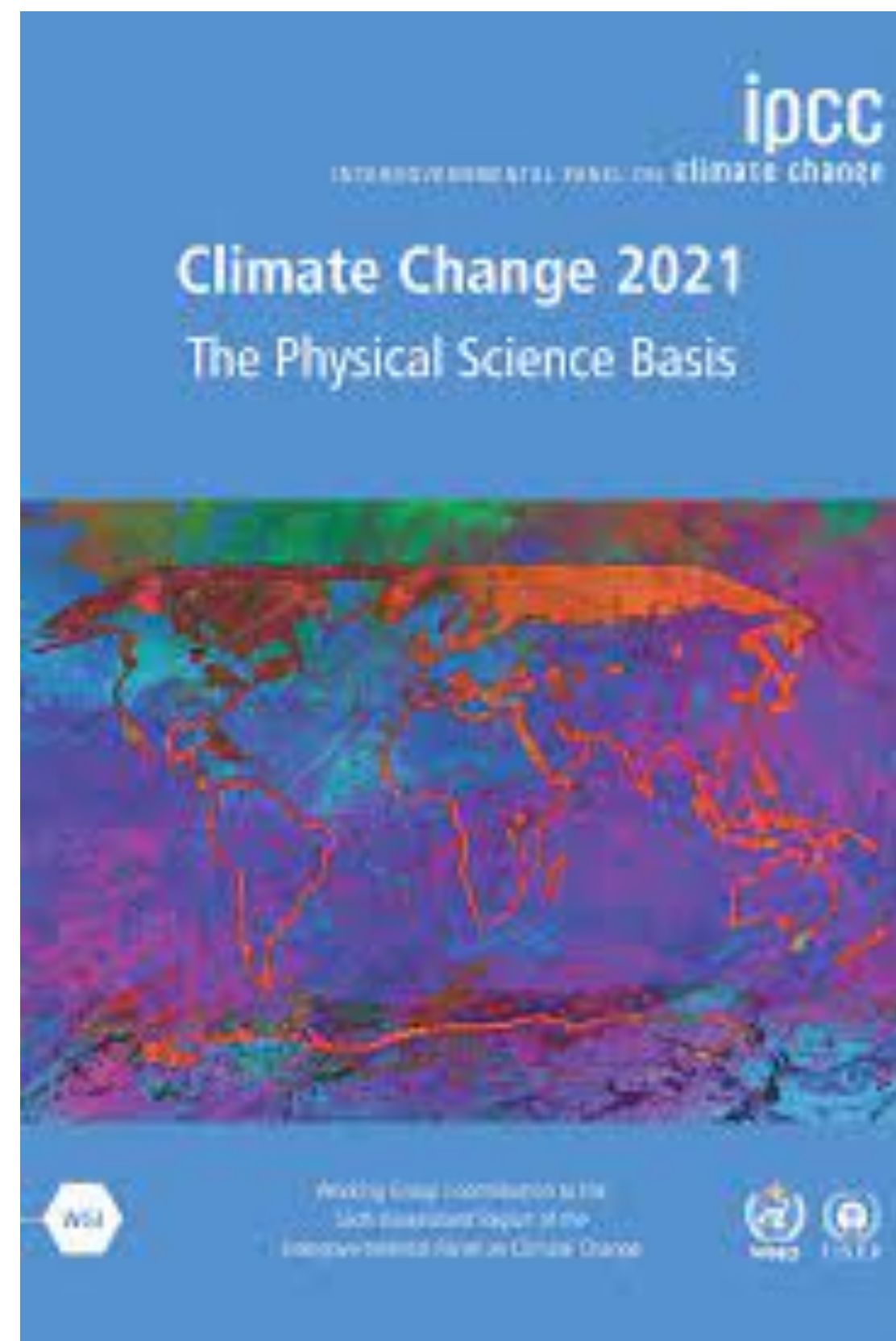
Romain Lacombe <rlacombe@stanford.edu> | August 28, 2025

How to calibrate LLM confidence?

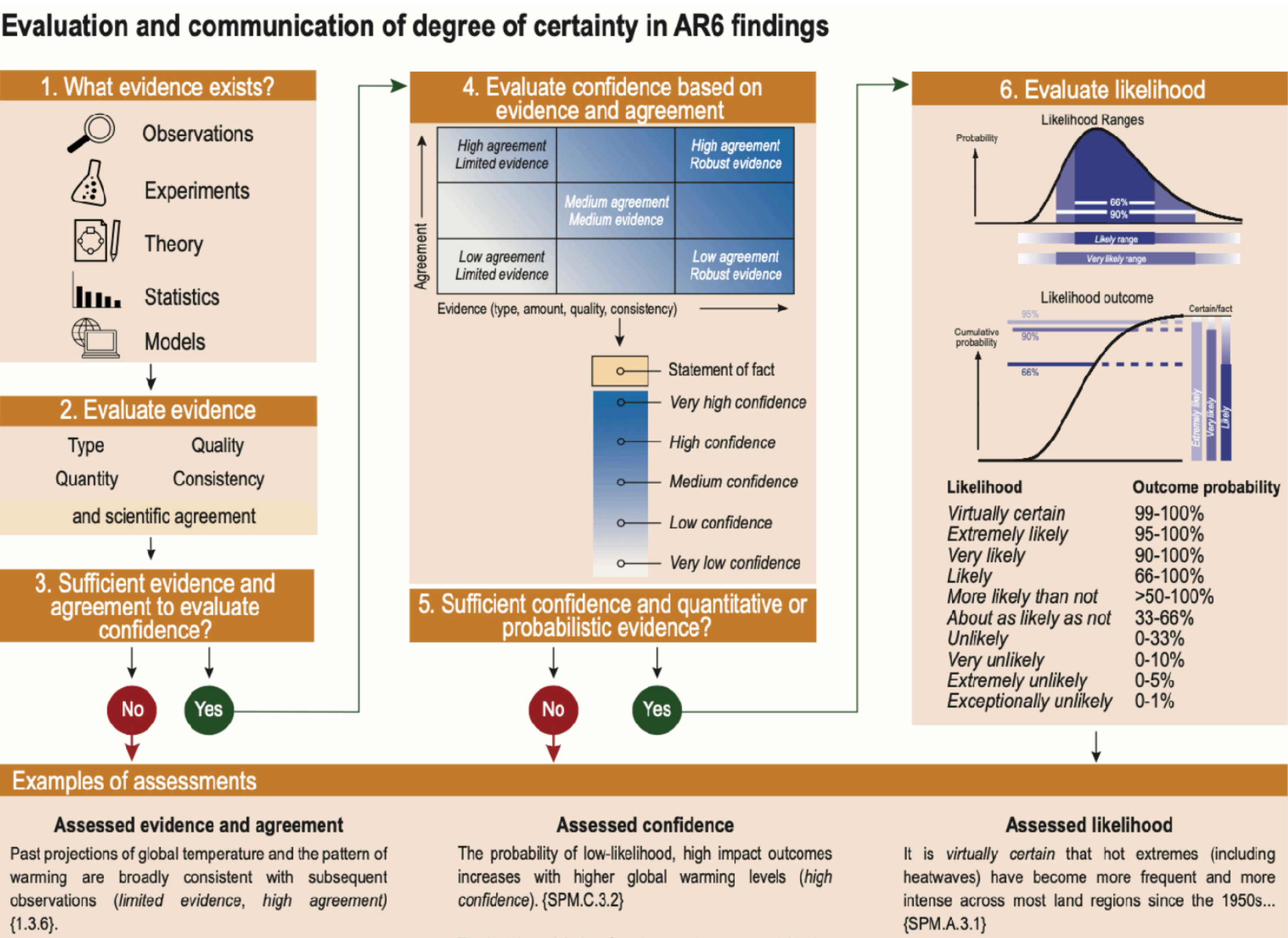
Against which ground truth?



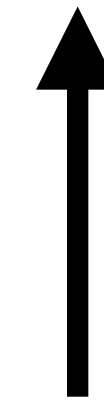
IPCC Assessment Reports on Climate Change



IPCC Guidelines to Authors on Confidence and Uncertainty Communication (AR6)



“X is caused by climate change (_____ confidence)”



low

medium

high

very high

ClimateX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?

Romain Lacombe, Kerrie Wu, Eddie Dilworth

NeurIPS 2023 Tackling Climate Change
with Machine Learning Workshop

SCAN ME



arXiv:2311.17107v1 [cs.LG] 28 Nov 2023

CLIMATEX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?

Romain Lacombe
Stanford University
rlacombe@stanford.edu

Kerrie Wu
Stanford University
kerriewu@stanford.edu

Eddie Dilworth
Stanford University
edjd@stanford.edu

Abstract

Evaluating the accuracy of outputs generated by Large Language Models (LLMs) is especially important in the climate science and policy domain. We introduce the Expert Confidence in Climate Statements (CLIMATEX) dataset, a novel, curated, expert-labeled dataset consisting of 8094 climate statements collected from the latest Intergovernmental Panel on Climate Change (IPCC) reports, labeled with their associated confidence levels. Using this dataset, we show that recent LLMs can classify human expert confidence in climate-related statements, especially in a few-shot learning setting, but with limited (up to 47%) accuracy. Overall, models exhibit consistent and significant over-confidence on low and medium confidence statements. We highlight implications of our results for climate communication, LLMs evaluation strategies, and the use of LLMs in information retrieval systems.

1 Introduction

The wide deployment of Large Language Models (LLMs) as question-answering tools calls for nuanced evaluation of their outputs across knowledge domains. This is especially important in climate science, where the quality of the information sources shaping public opinion, and ultimately public policy, could determine whether the world succeeds or fails in tackling climate change.

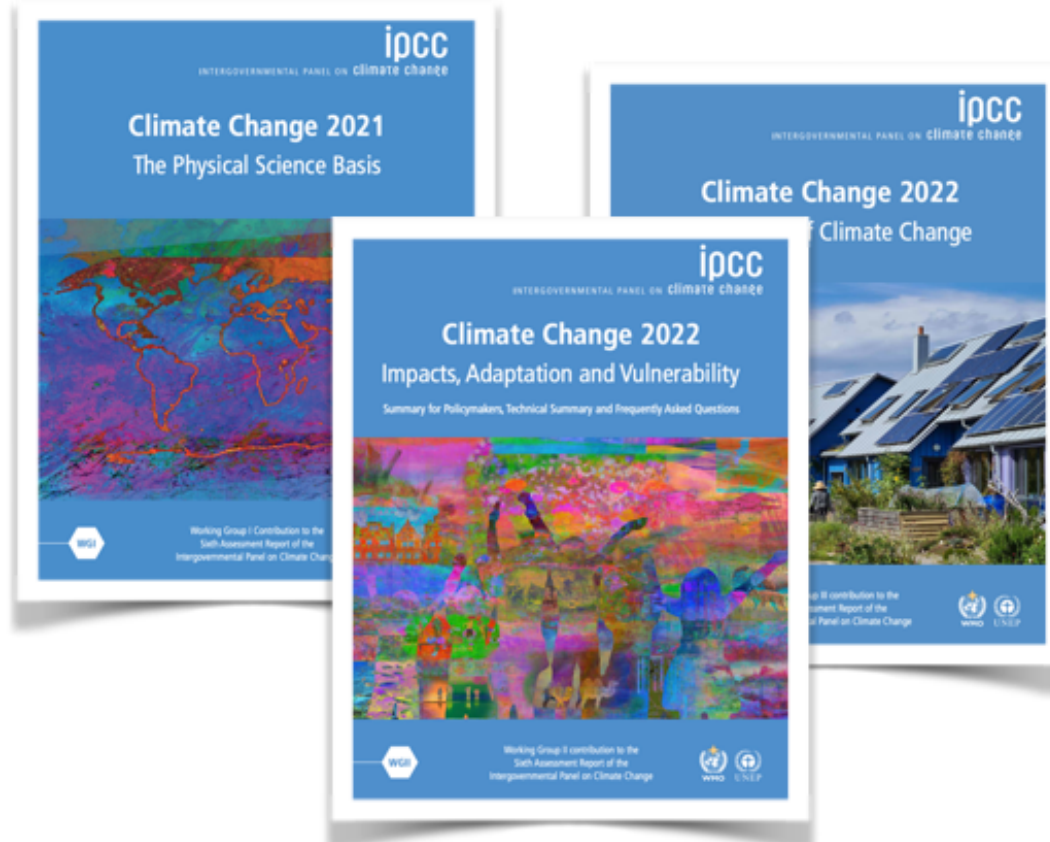
This paper aims to evaluate the reliability of LLM outputs in the climate science and policy domain. We introduce the **Expert Confidence in Climate Statements (CLIMATEX) dataset** [10], a novel, curated, expert-labeled, natural language dataset of 8094 statements sourced from the three most recent Intergovernmental Panel on Climate Change Assessment Reports (IPCC AR6) — along with their associated confidence levels (low, medium, high, or very high) that were assessed by climate scientists based on the quantity and quality of available evidence and agreement among their peers. CLIMATEX is available on HuggingFace and source code for experiments is available on Github.

We use this dataset to evaluate how accurately recent LLMs assess the confidence which human experts associate with climate science statements. Although OpenAI's GPT-3.5-turbo and GPT-4 assess the true confidence level with better-than-random accuracy and higher performance than non-expert humans, even in a zero-shot setting, they, and other models we tested, consistently overstate the certainty level associated with low and medium confidence labeled statements.

With LLMs poised to become increasingly significant sources of public information, the reliability of their outputs in the climate domain is critical for avoiding misinformation and garnering support for effective climate policy. We hope the CLIMATEX dataset provides a valuable tool for benchmarking the trustworthiness of LLM outputs in the climate domain, highlights the need for further work in this area, and aids efforts to develop models that accurately convey climate knowledge.

ClimateX Dataset

IPCC Reports AR6 WGI/II/III



8094 statements with confidence labels

Statement

`<sentence>`
(`{low|medium|high`
`|very high}`
`confidence)`

Label

Test set Manual clean up

300 statements
Randomly selected
Human expert labeled

Remove all citations
`<* et al., 20??>`

Expand all acronyms
in test set

Train set Automated clean up

7794 statements
(Remainder of the
8094 statements)

Remove found citations
`<* et al., 20??>`

Expand 66 acronyms
found in test set

Statement: “X is caused by climate change”
Confidence? _____



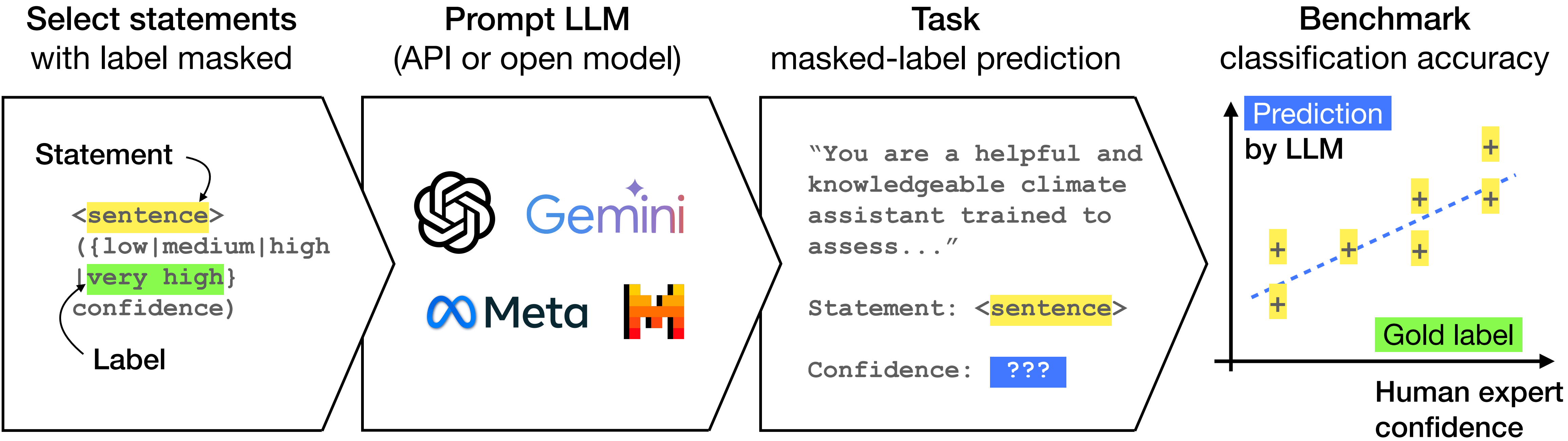
low

medium

high

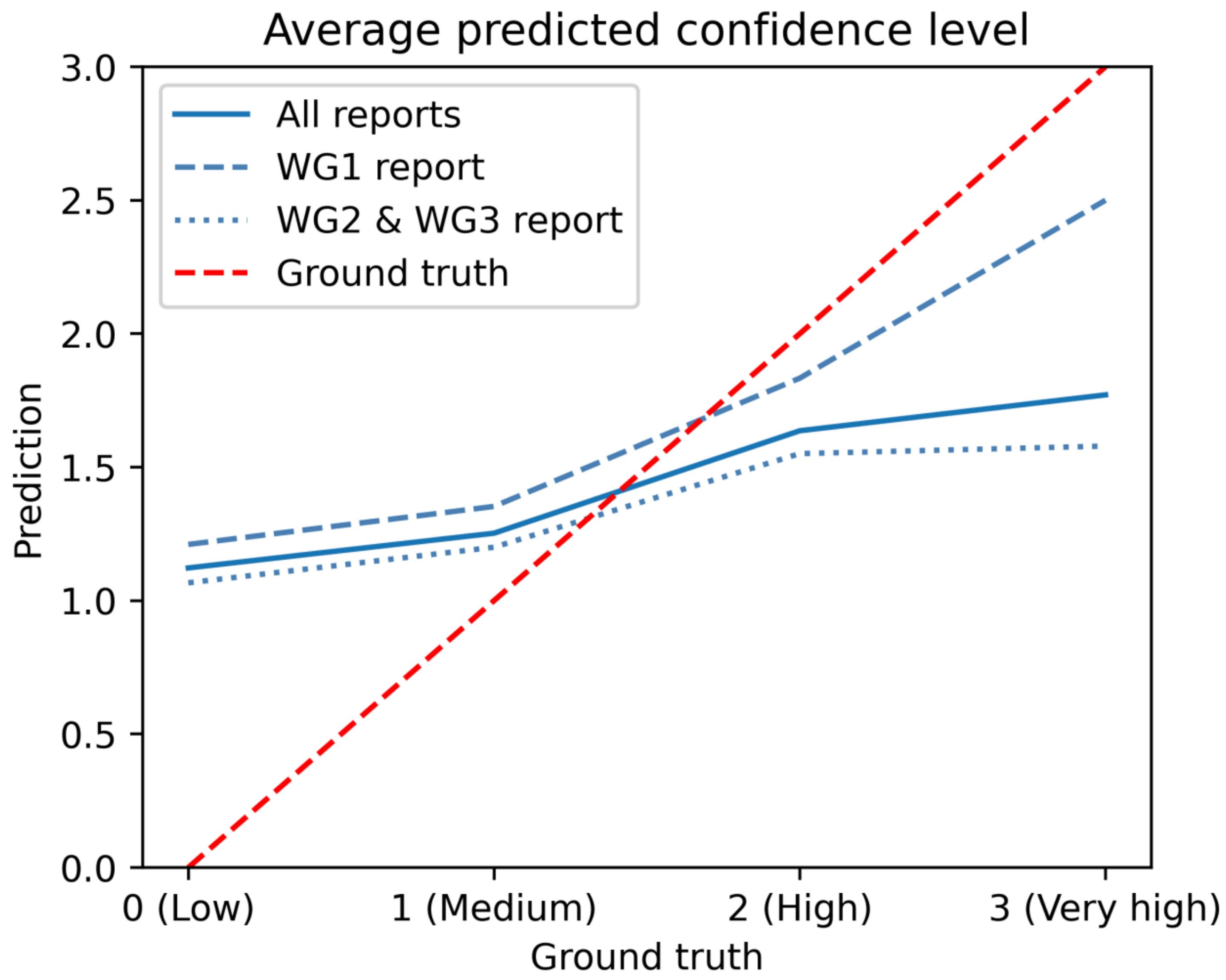
very high

ClimateX Benchmark: masked confidence label prediction



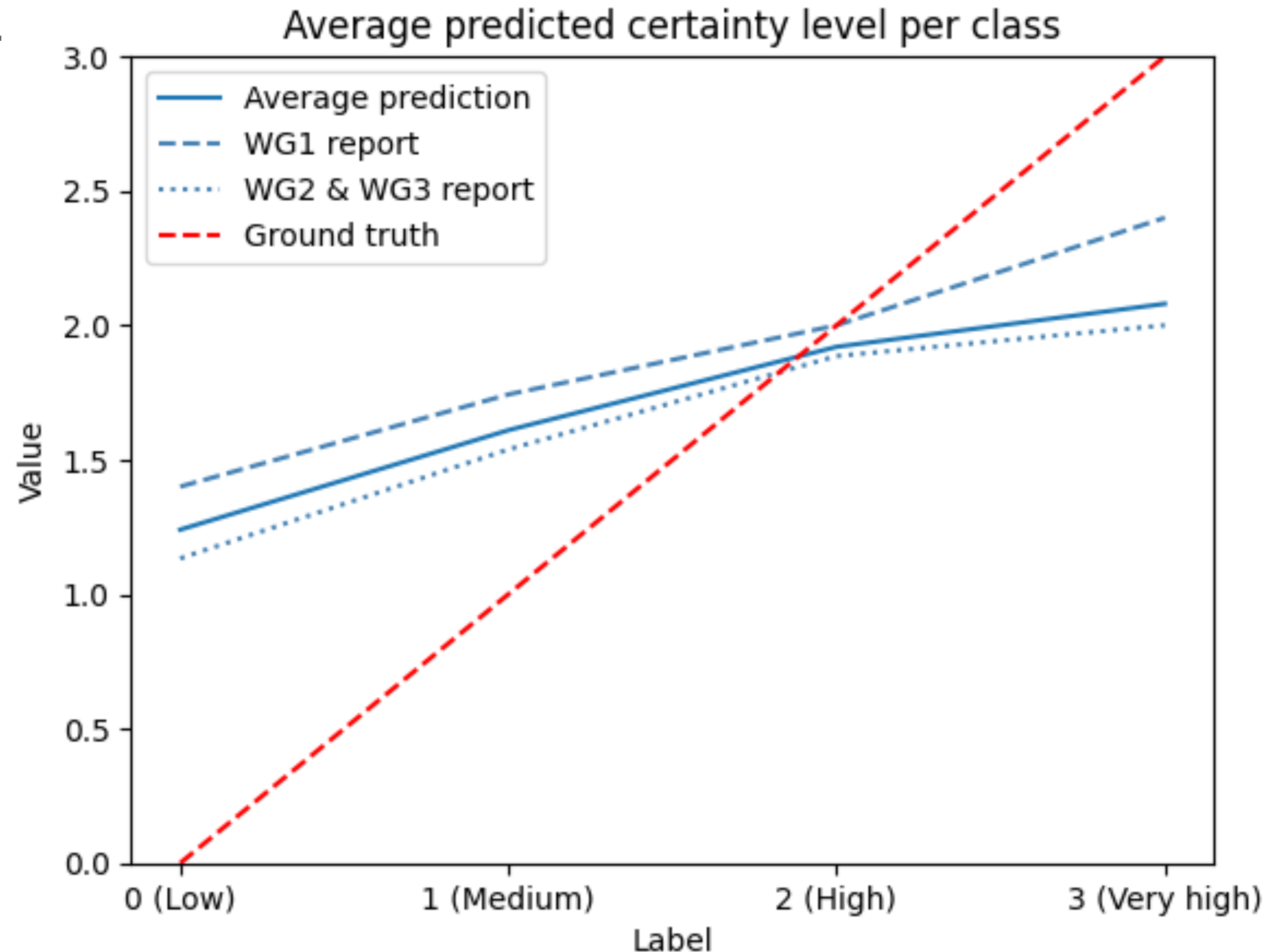
GPT-3.5

June 2023



Gemini Pro 1.0

June 2024



ClimateX Results | June 2024

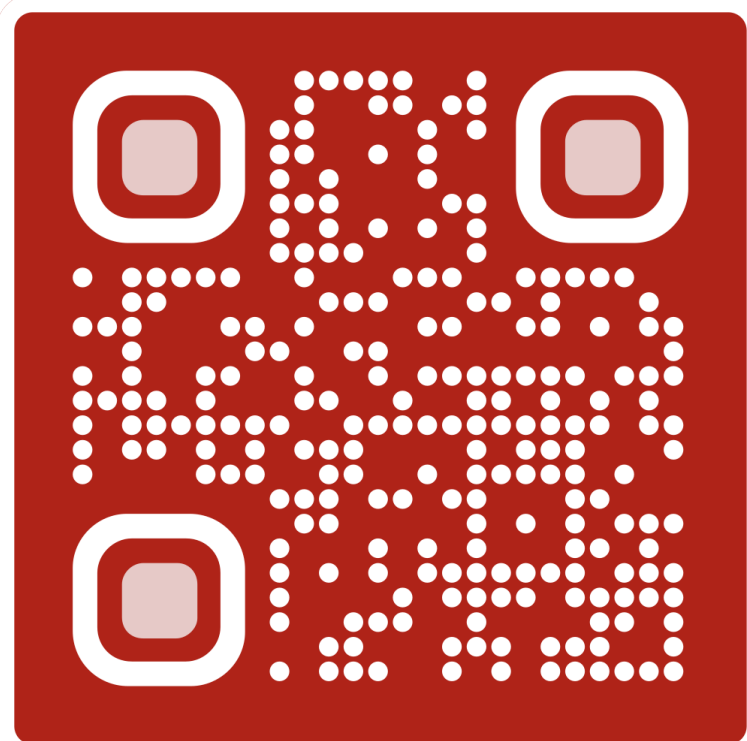
Model	Accuracy	Slope	Bias	Parameters
LLM APIs				
Google Gemini Pro	45.0% ± 0.0	0.285 ± 0.000	0.230 ± 0.000	Unkown
OpenAI GPT-4o	44.0% ± 1.1	0.350 ± 0.011	0.283 ± 0.007	Unkown
OpenAI GPT-4	42.4% ± 0.5	0.233 ± 0.007	0.197 ± 0.007	Unkown
OpenAI GPT-3.5 Turbo	39.7% ± 0.6	0.153 ± 0.008	0.226 ± 0.010	Unkown
Open-Source LLMs				
Meta Llama 3 8B Chat	41.1% ± 0.3	0.120 ± 0.005	-0.001 ± 0.006	8B
Mixtral-8x22B Instruct v0.1	38.1% ± 0.3	0.360 ± 0.004	0.418 ± 0.002	8 \times 22B
Meta Llama 3 70B Chat	36.2% ± 0.3	0.239 ± 0.003	0.444 ± 0.010	70B
Mixtral-8x7B Instruct v0.1	35.9% ± 0.3	0.187 ± 0.011	0.303 ± 0.005	8 \times 7B
Mistral 7B Instruct v0.3	35.0% ± 0.0	0.235 ± 0.000	0.423 ± 0.000	7B
Google Gemma Instruct 2B	33.9% ± 0.0	0.062 ± 0.000	0.010 ± 0.000	2B
Google Gemma Instruct 7B	33.4% ± 0.3	0.049 ± 0.009	0.305 ± 0.005	7B
Baselines				
RoBERTa	53.7%			
Non-expert humans	36.2%			

**Do reasoning models have
better confidence calibration?**

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

Romain Lacombe, Kerrie Wu,
Eddie Dilworth

ICML 2025 Reliable and Responsible
Foundation Models workshop



arXiv:2508.15050v1 [cs.AI] 20 Aug 2025

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

Romain Lacombe¹ Kerrie Wu¹ Eddie Dilworth¹

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.

¹Stanford 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.

ClimateX Results

May 2025

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

Tool Use: Search

Gemini 2.5

Model	Accuracy	Cohen's κ	Bias
Search-Augmented Models			
Google Gemini 2.5 Pro with Search	89.3%	85.7%	+0.030
Google Gemini 2.5 Flash with Search	88.3%	84.4%	+0.097
Reasoning Models			
Google Gemini 2.5 Pro	48.7%	31.6%	+0.066
Google Gemini 2.5 Pro – Bulk processing	45.3%	27.1%	+0.353
Google Gemini 2.5 Flash – Best thinking budget	45.0%	26.7%	+0.265
OpenAI o3 – Program synthesis	40.7%	20.9%	+0.167

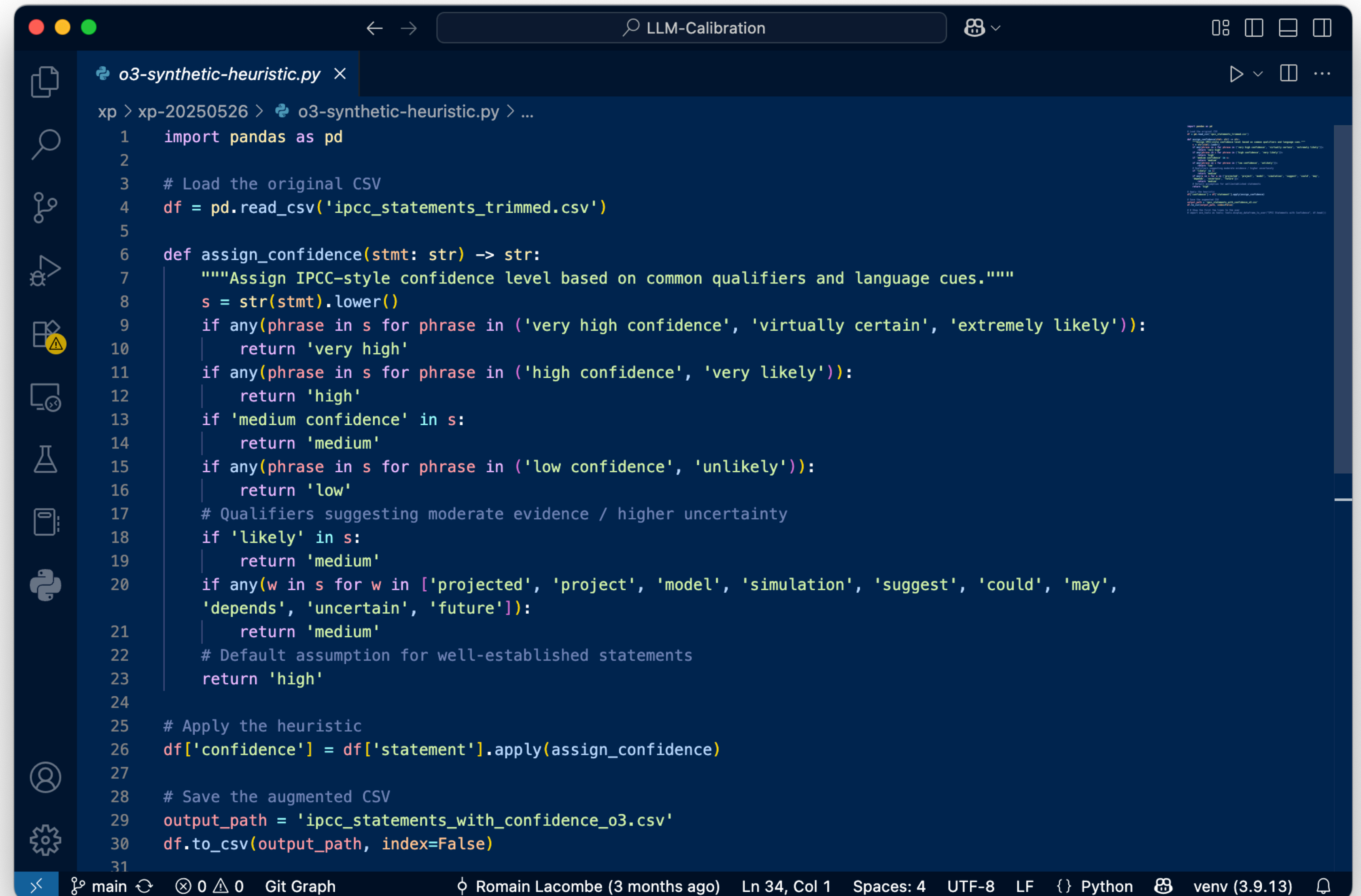
Long context windows enable bulk processing

Gemini 2.5 Pro

Model	Accuracy	Cohen's κ	Bias
Search-Augmented Models			
Google Gemini 2.5 Pro with Search	89.3%	85.7%	+0.030
Google Gemini 2.5 Flash with Search	88.3%	84.4%	+0.097
Reasoning Models			
Google Gemini 2.5 Pro	48.7%	31.6%	+0.066
Google Gemini 2.5 Pro – Bulk processing	45.3%	27.1%	+0.353
Google Gemini 2.5 Flash – Best thinking budget	45.0%	26.7%	+0.265
OpenAI o3 – Program synthesis	40.7%	20.9%	+0.167

Program synthesis

OpenAI o3



```
o3-synthetic-heuristic.py x
xp > xp-20250526 > o3-synthetic-heuristic.py > ...
1 import pandas as pd
2
3 # Load the original CSV
4 df = pd.read_csv('ipcc_statements_trimmed.csv')
5
6 def assign_confidence(stmt: str) -> str:
7     """Assign IPCC-style confidence level based on common qualifiers and language cues."""
8     s = str(stmt).lower()
9     if any(phrase in s for phrase in ('very high confidence', 'virtually certain', 'extremely likely')):
10         return 'very high'
11     if any(phrase in s for phrase in ('high confidence', 'very likely')):
12         return 'high'
13     if 'medium confidence' in s:
14         return 'medium'
15     if any(phrase in s for phrase in ('low confidence', 'unlikely')):
16         return 'low'
17     # Qualifiers suggesting moderate evidence / higher uncertainty
18     if 'likely' in s:
19         return 'medium'
20     if any(w in s for w in ['projected', 'project', 'model', 'simulation', 'suggest', 'could', 'may',
21                             'depends', 'uncertain', 'future']):
22         return 'medium'
23     # Default assumption for well-established statements
24     return 'high'
25
26 # Apply the heuristic
27 df['confidence'] = df['statement'].apply(assign_confidence)
28
29 # Save the augmented CSV
30 output_path = 'ipcc_statements_with_confidence_o3.csv'
31 df.to_csv(output_path, index=False)
```

LLM-Calibration

main 0 0 Git Graph Romain Lacombe (3 months ago) Ln 34, Col 1 Spaces: 4 UTF-8 LF {} Python venv (3.9.13)

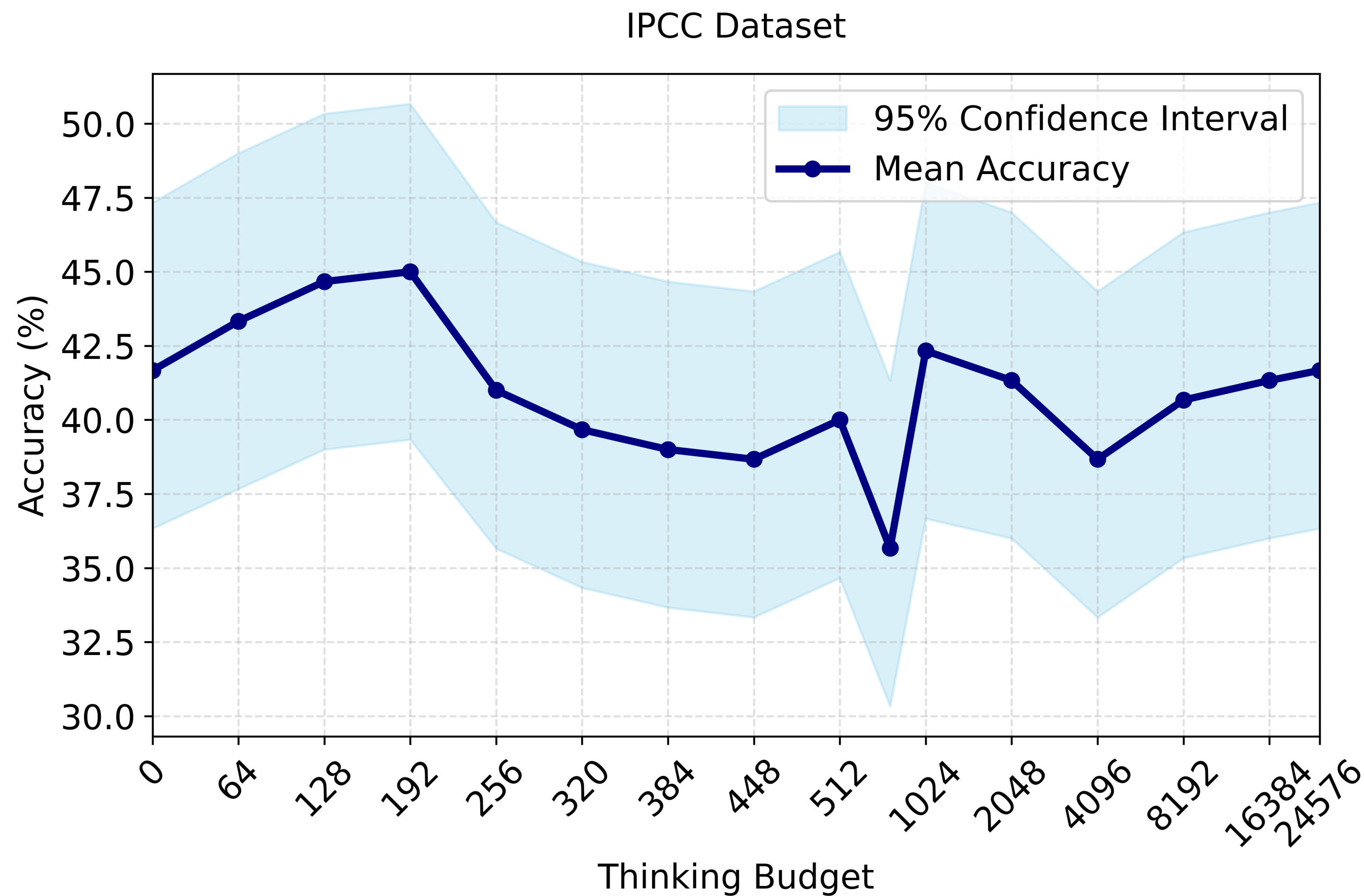
What's the optimal thinking budget?

Gemini 2.5 Flash

Model	Accuracy	Cohen's κ	Bias
Search-Augmented Models			
Google Gemini 2.5 Pro with Search	89.3%	85.7%	+0.030
Google Gemini 2.5 Flash with Search	88.3%	84.4%	+0.097
Reasoning Models			
Google Gemini 2.5 Pro	48.7%	31.6%	+0.066
Google Gemini 2.5 Pro – Bulk processing	45.3%	27.1%	+0.353
Google Gemini 2.5 Flash – Best thinking budget	45.0%	26.7%	+0.265
OpenAI o3 – Program synthesis	40.7%	20.9%	+0.167

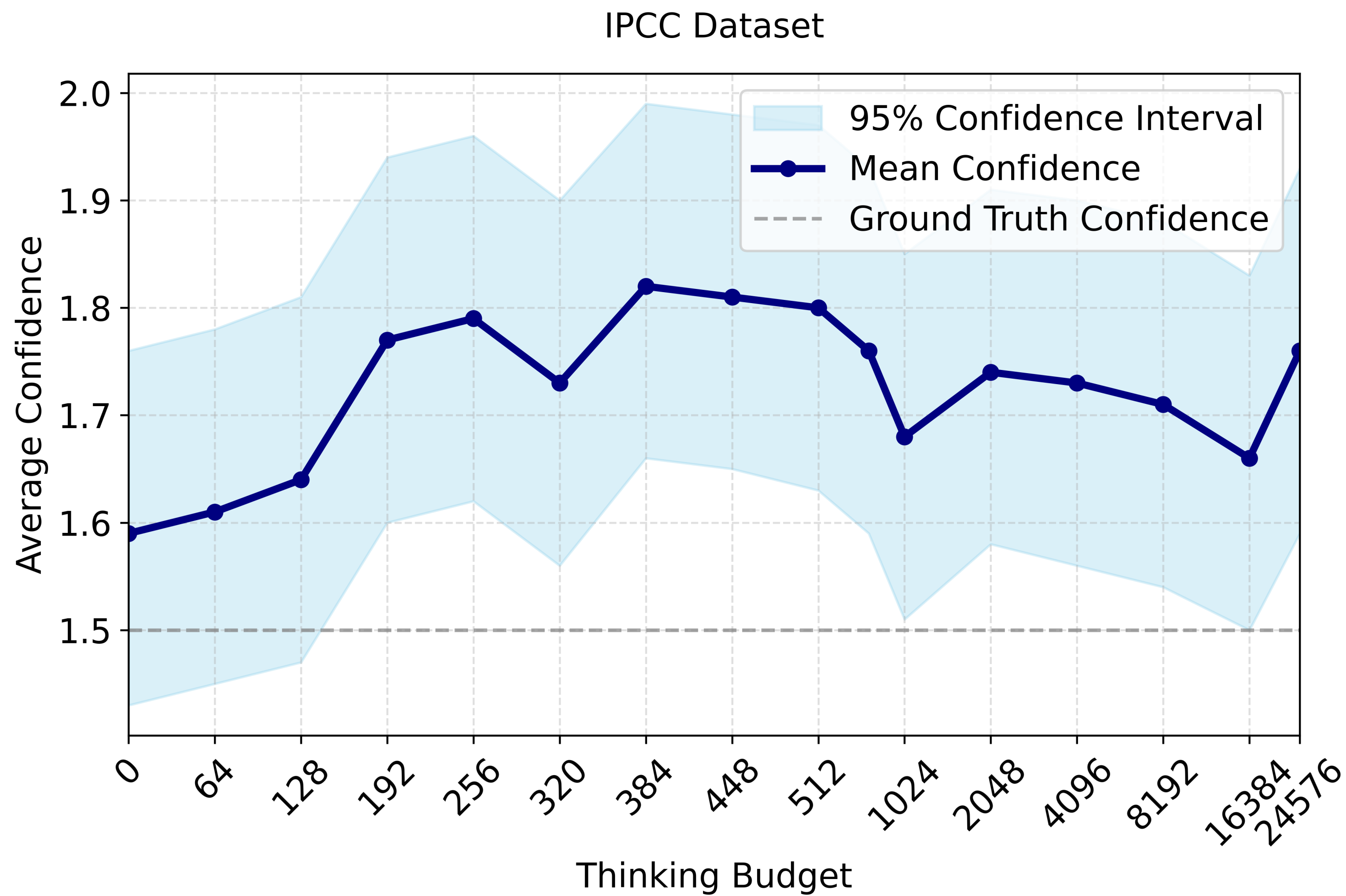
Thinking Budget vs Accuracy (0%-100%)

Gemini Flash 2.5



Thinking Budget vs Confidence (0.0-3.0)

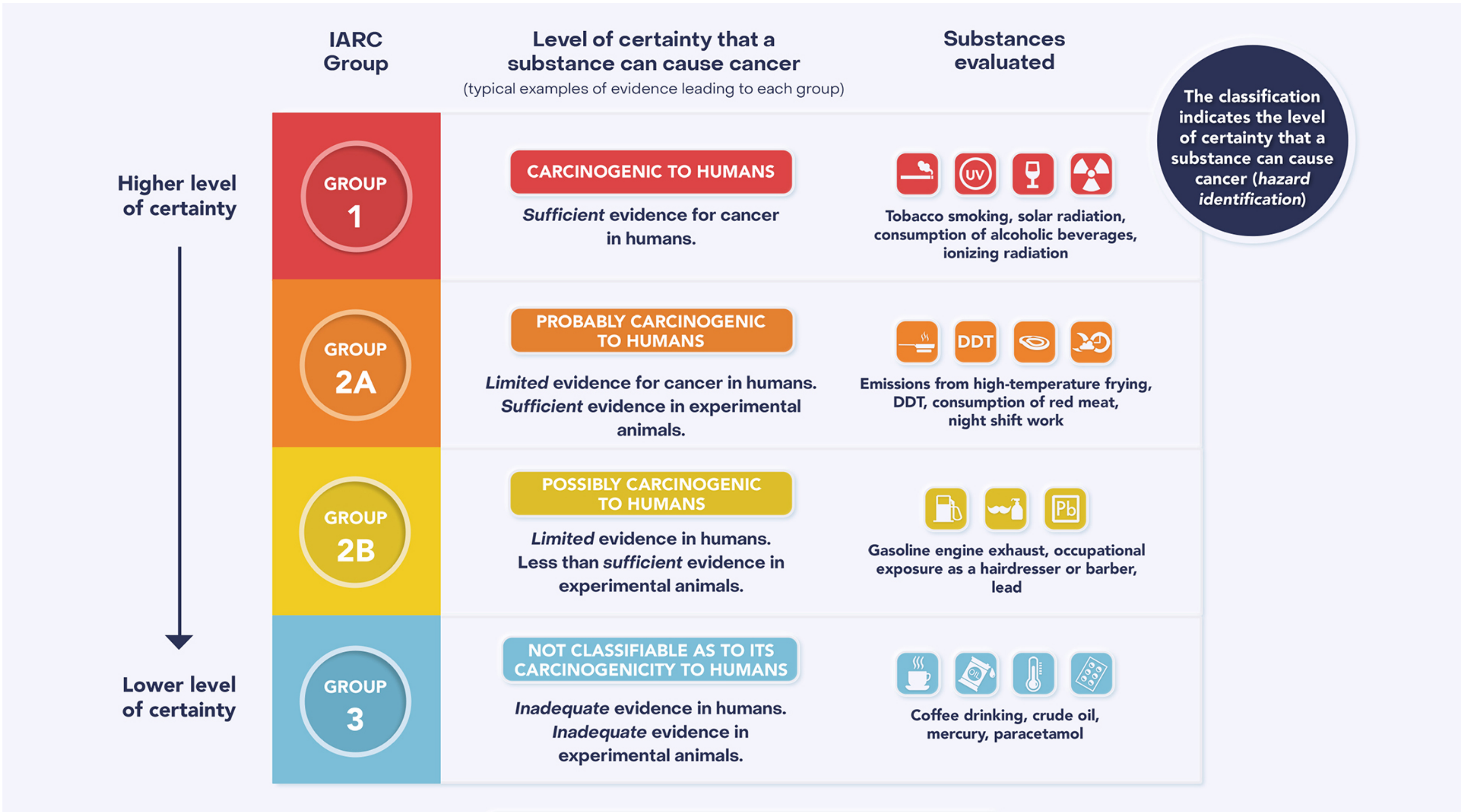
Gemini Flash 2.5



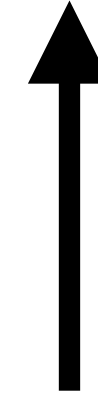
**Does this generalize
beyond climate science?**

Public Health & Oncology

WHO IARC Carcinogenicity Monographs



“Substance X is _____ carcinogenic”



not classifiable as

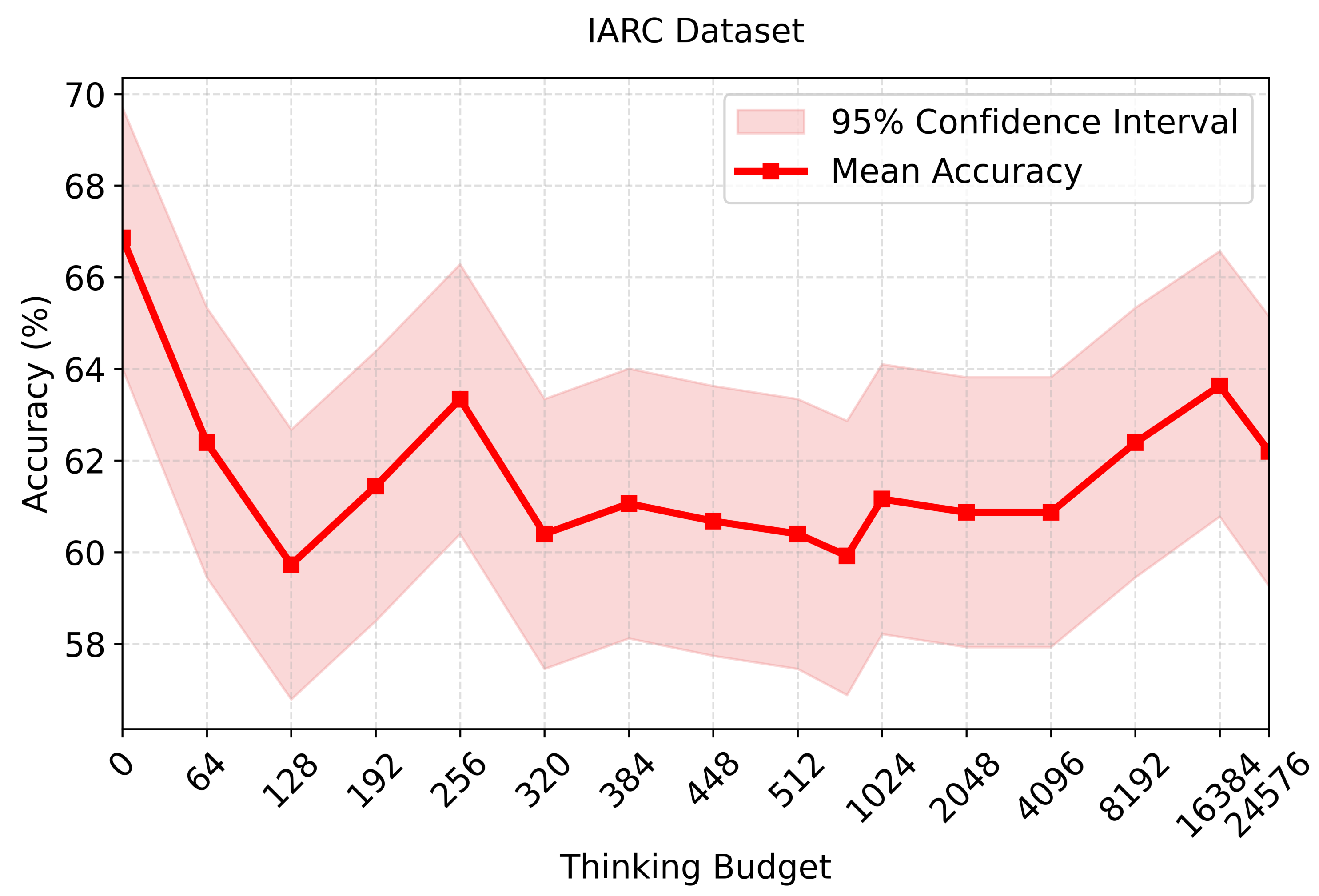
possibly

probably

known to be

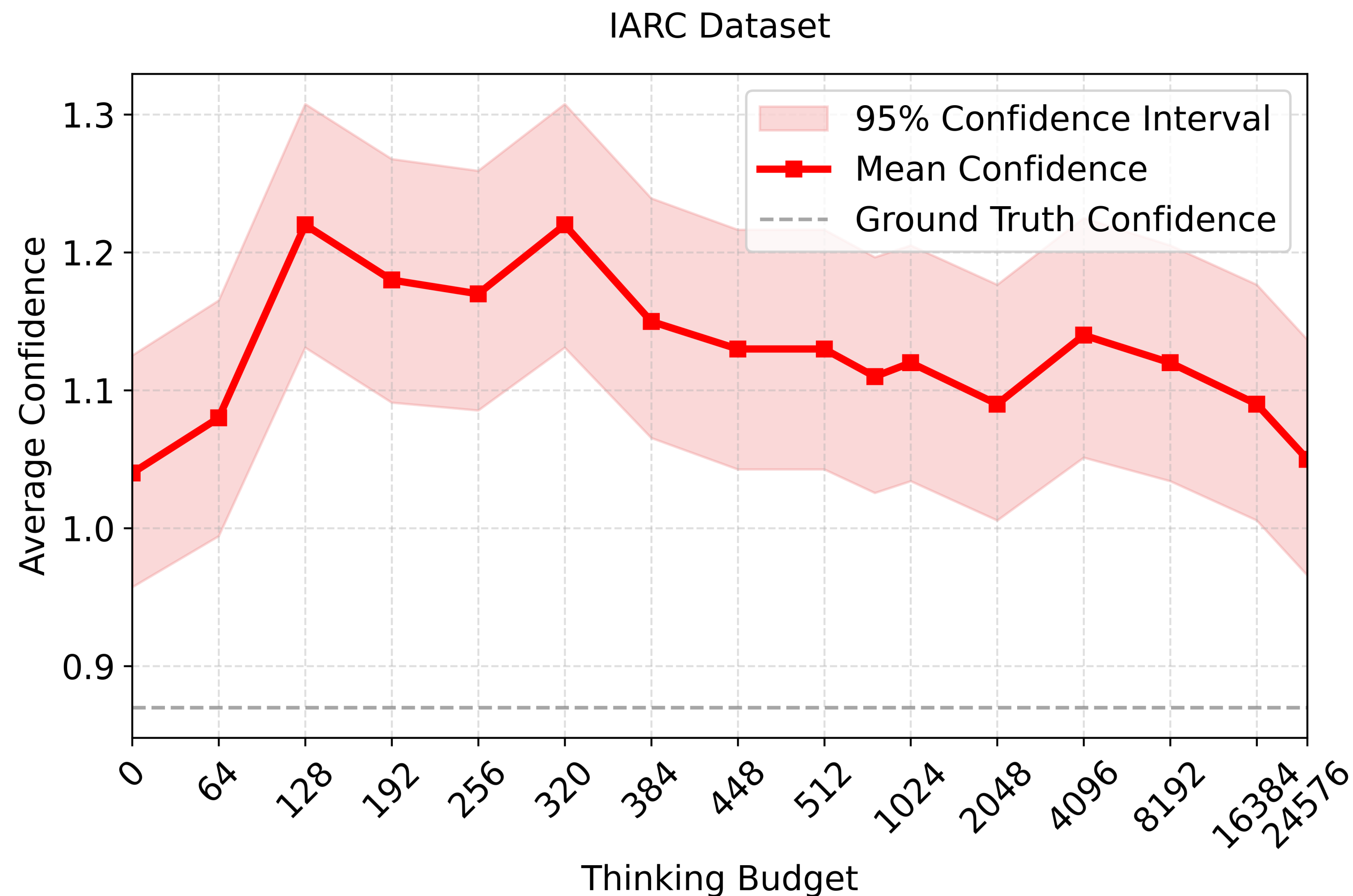
Thinking Budget vs Accuracy (0%-100%)

Gemini Flash 2.5



Thinking Budget vs Confidence (0.0-3.0)

Gemini Flash 2.5



**Take aways and
future work?**

Increasing thinking budget through test-time scaling *impairs* the confidence calibration of reasoning models.

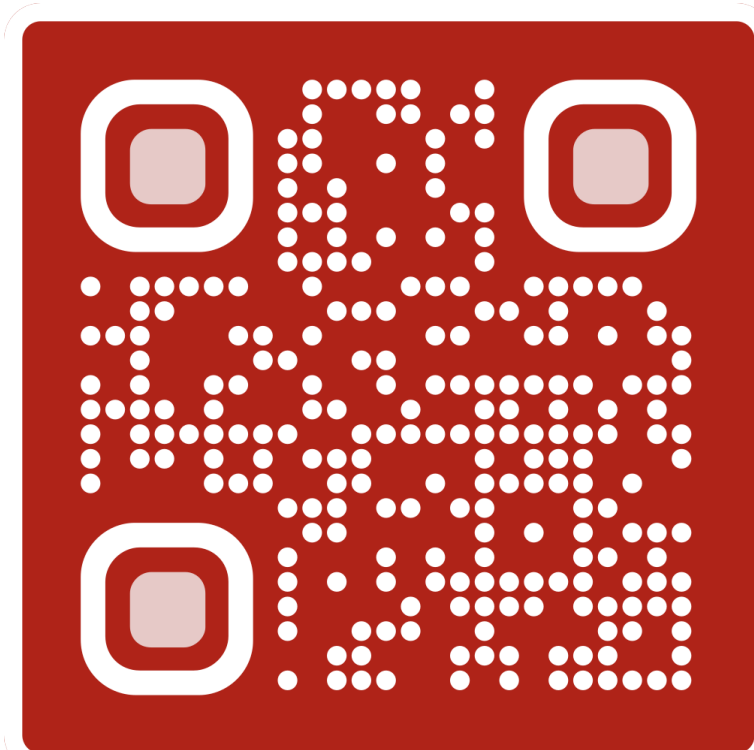
LLM grounding is bottlenecked by [access](#) to or recall of relevant evidence rather than by reasoning capacity.

Emerging capability: program-synthesis fallback, when models unable to perform the task directly attempt to generate algorithmic solutions.

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

Romain Lacombe, Kerrie Wu,
Eddie Dilworth

ICML 2025 Reliable and Responsible
Foundation Models workshop



arXiv:2508.15050v1 [cs.AI] 20 Aug 2025

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

Romain Lacombe¹ Kerrie Wu¹ Eddie Dilworth¹

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.

¹Stanford 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.

Thank you!



Romain Lacombe
Stanford ChemE



Kerrie Wu
Stanford CS



Eddie Dilworth
Stanford CS



Chris Potts
Stanford NLP



ClimateChange AI
NeurIPS Workshop

Stanford



Thank you!

Questions? :)

Romain Lacombe <rlacombe@stanfordalumni.org> | August 28, 2025