Stanford

Stanford Data Science

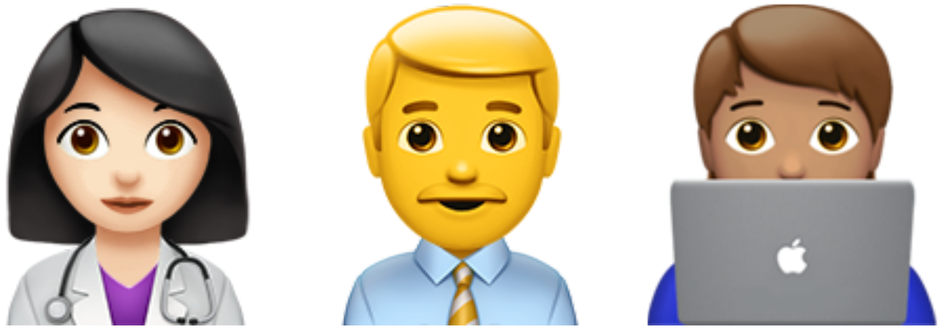
ClimateX

Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?

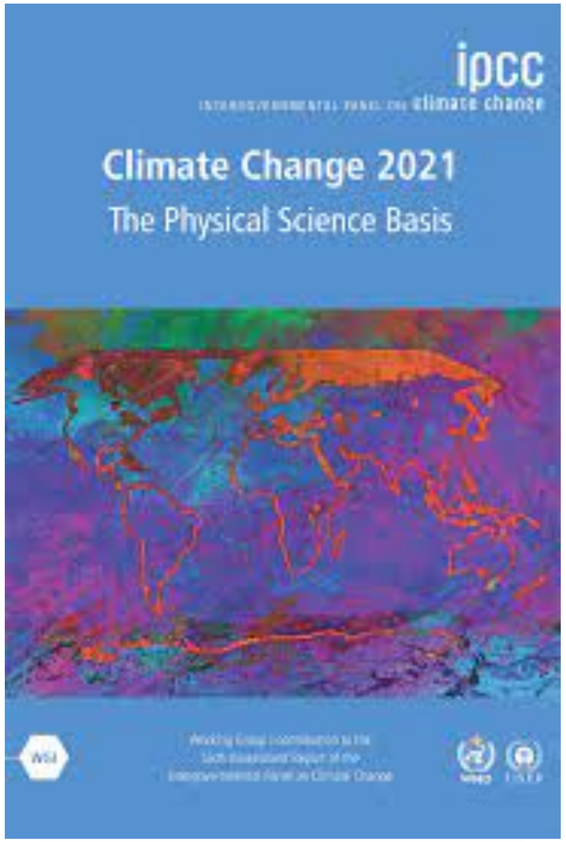
Romain Lacombe <rlacombe@stanford.edu> | April 10, 2025

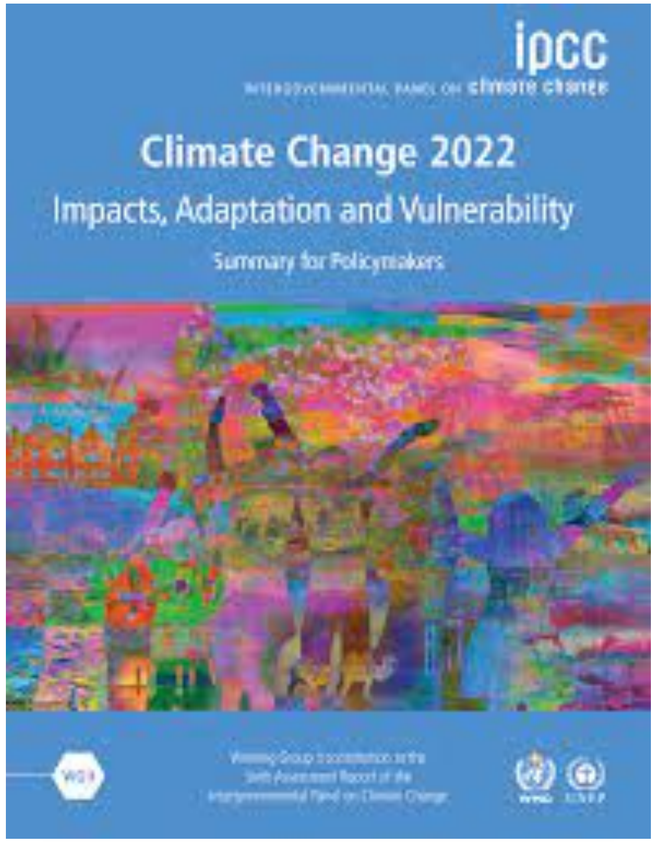
How to calibrate LLM confidence?

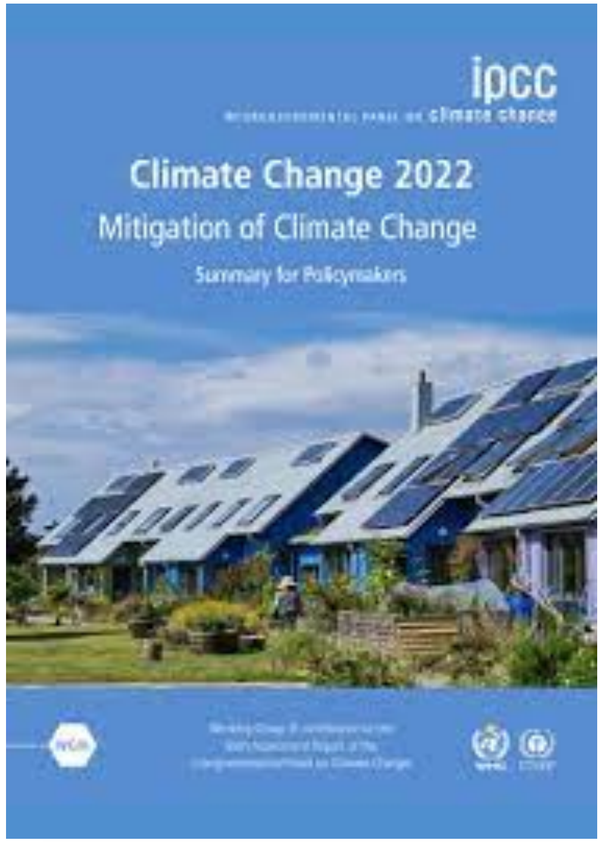
Against which ground truth?



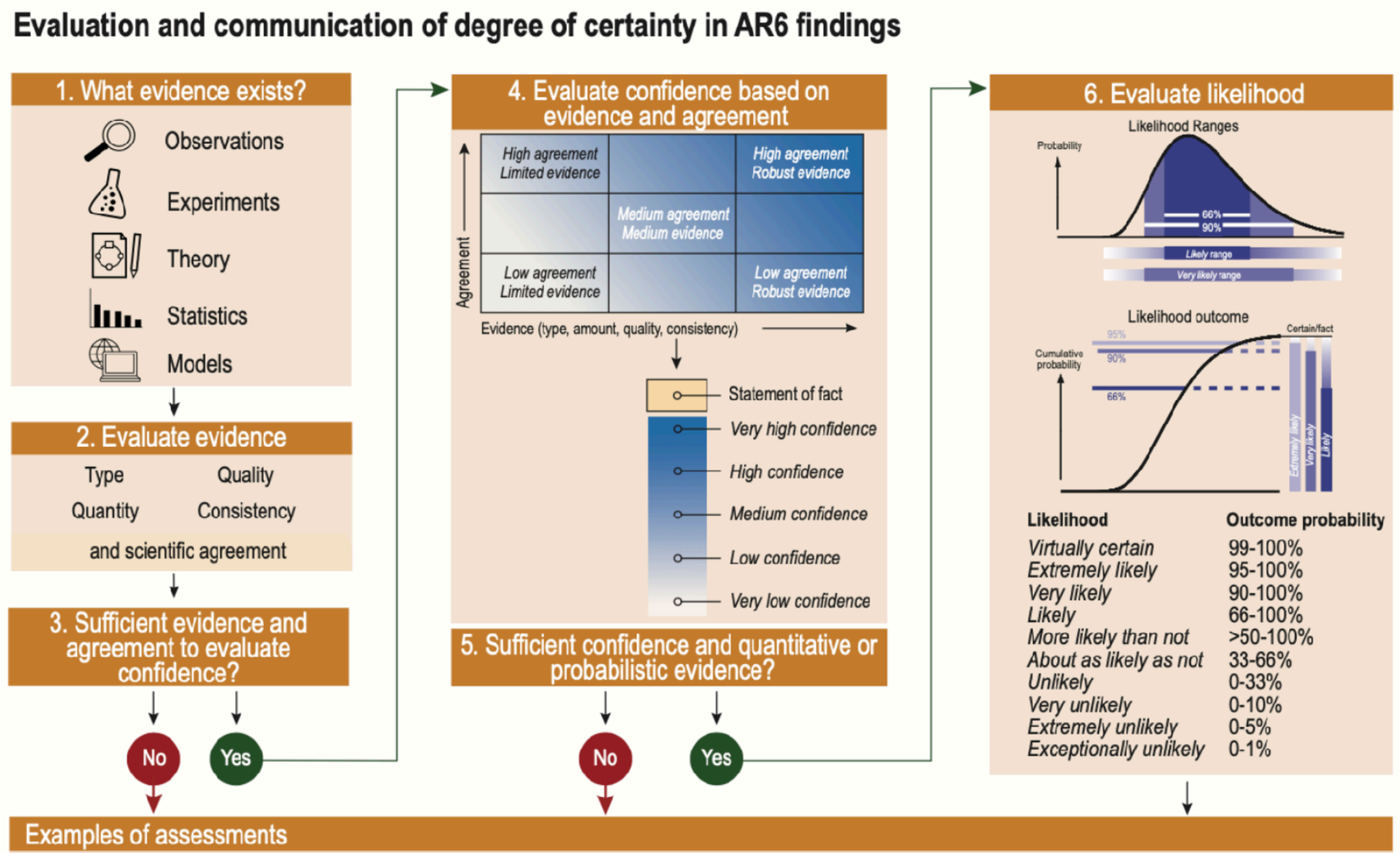
IPCC Assessment Reports on Climate Change



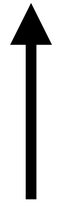




IPCC Guidelines to Authors on Confidence and Uncertainty Communication (AR6)



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



low

medium

high

very high

ClimateX 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? _____

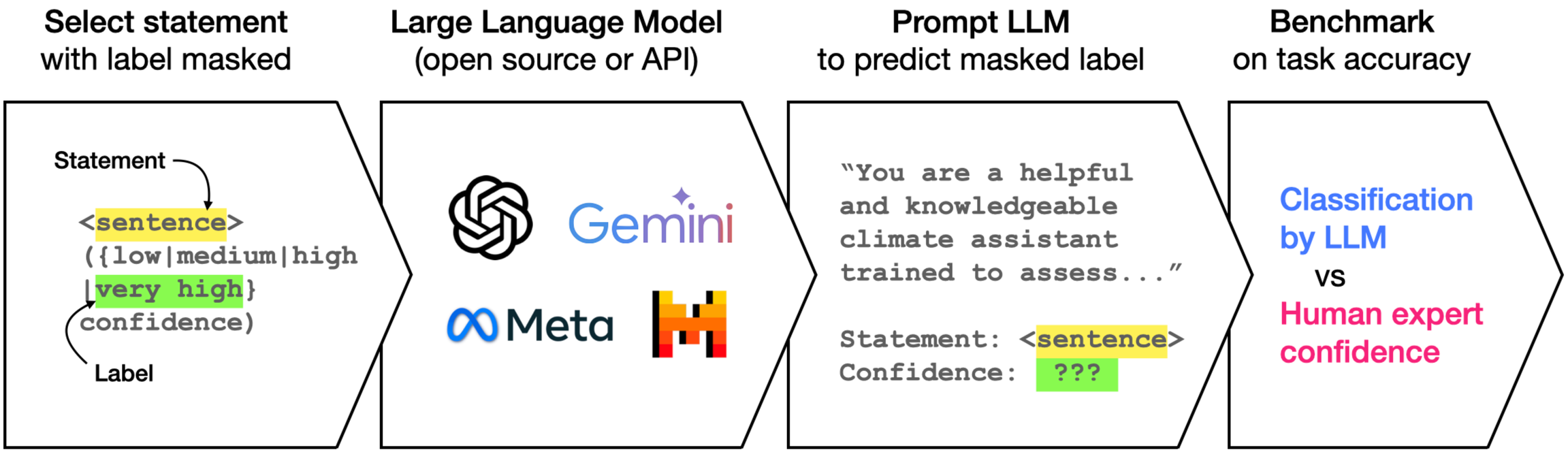
low

medium

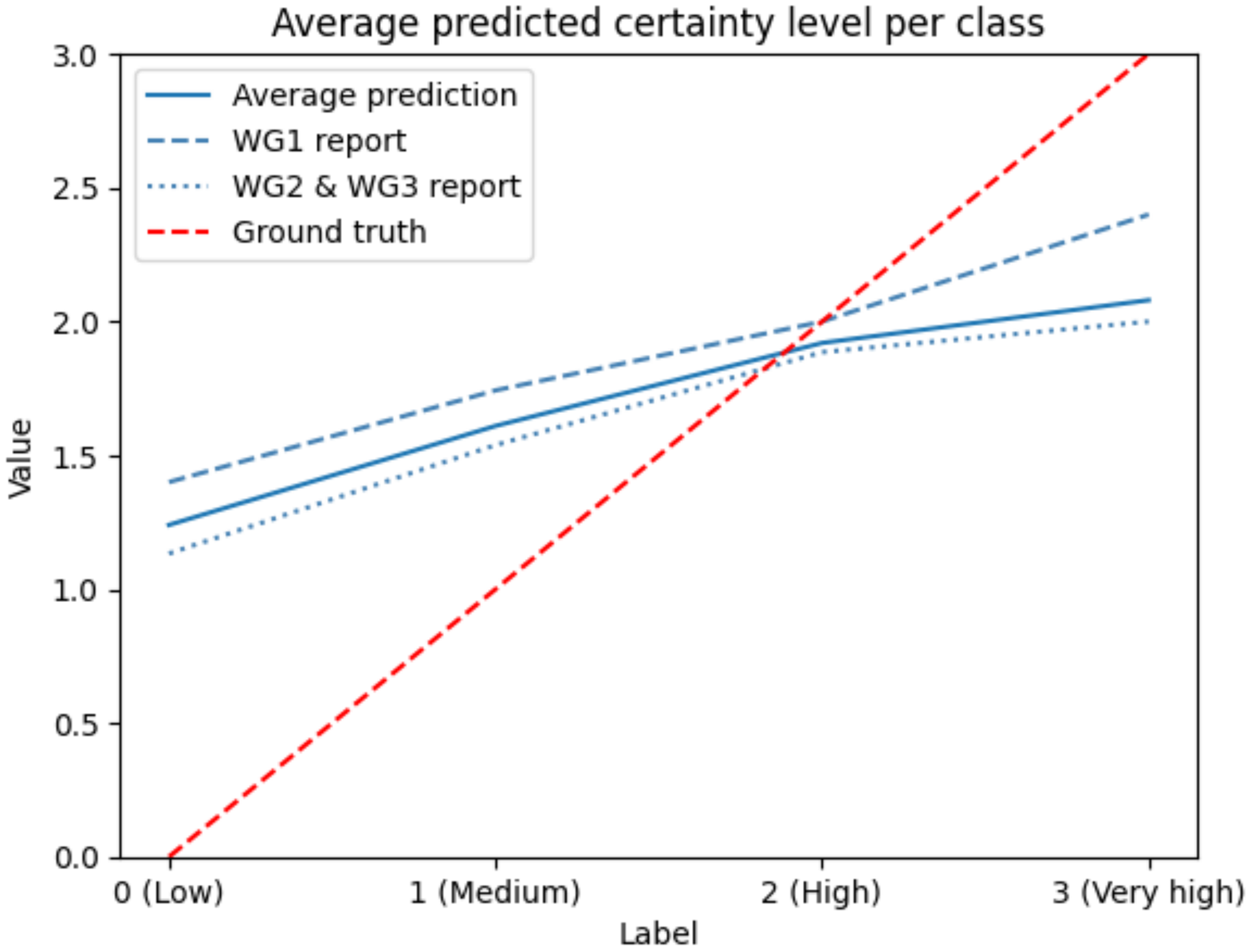
high

very high

ClimateX Benchmark



Gemini Pro 1.0

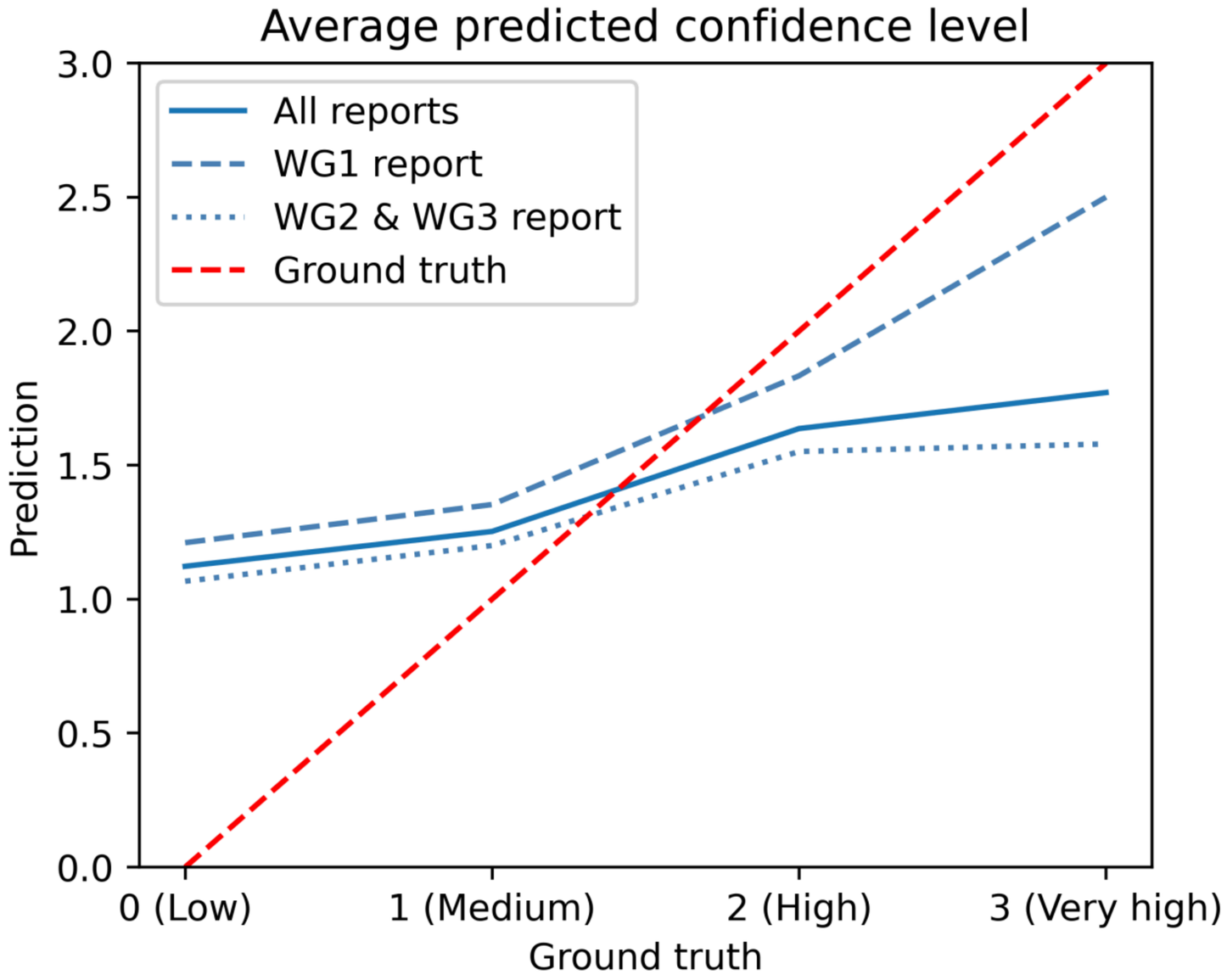


ClimateX Results | December 2024

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	8×22B
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	8×7B
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%			

Limitations

GPT-3.5


Average predicted confidence level
All reports
WG1 report
WG2 & WG3 report
Ground truth
Prediction
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0 (Low)
1 (Medium)
2 (High)
3 (Very high)
Ground truth

Future work

Thank you!

Romain Lacombe
Stanford ChemE

Kerrie Wu
Stanford CS

Eddie Dilworth
Stanford CS

Stanford | NLP

Chris Potts
Stanford NLP

Climate Change AI

ClimateChange AI
NeurIPS Workshop