Group Meeting

Yunlong Pan

2024-05-27

Outline

- ClimateX
- Our ChatIAMs
- Result Comparison
- Few-shots(Not random)

ClimateX

Source:

1. Paper: <<ClimateX: Do LLMs Accurately Assess Human Expert Confidence in Climate Statements?>> https://arxiv.org/abs/2311.17107
2. Github Code: https://github.com/rlacombe/ClimateX/tree/main
3. Dataset: https://huggingface.co/datasets/rlacombe/ClimateX

statement_idx int64 0–809 10%	report string · classes AR6_WGI 28.…	page_num int64 18–300 25%	sent_num int64 16–23 24%	statement string · lengths 441–869 2.4%	confidence string · classes high 58.…	score int64 2 58.…	split string · classes train 96.…
0	AR6_WGI	20	22	Since 2011 (measurements reported in AR5), concentrations have continued to increase in the atmosphere, reaching annual averages of 410 parts per million (ppm) for carbon dioxide (CO 2), 1866 parts per billion (ppb) for methane (CH 4), and 332 ppb for nitrous oxide (N 2O) in 2019.6 Land and ocean have taken up a near-constant proportion (globally about 56% per year) of CO 2 emissions from human activities over the past six decades, with regional differences	high	2	train
1	AR6_WGI	21	8	Mid-latitude storm tracks have likely shifted poleward in both hemispheres since the 1980s, with marked…	medium	1	train
2	AR6_WGI	21	18	The average rate of sea level rise was 1.3 [0.6 to 2.1] mm yr−1 between 1901 and 1971, increasing to 1.9 [0.8 t…	high	2	train
3	AR6_WGI	24	2	Since 1750, increases in CO2 (47%) and CH4 (156%) concentrations far exceed – and increases in N2O (23%)…	very high	3	test
4	AR6_WGI	24	4	Temperatures during the most recent decade (2011–2020) exceed those of the most recent multi-century warm…	medium	1	train
5	AR6_WGI	24	5	Prior to that, the next most recent warm period was about 125,000 years ago, when the multi-century…	medium	1	train
6	AR6_WGI	24	7	Late summer Arctic sea ice area was smaller than at any time in at least the past 1000 years	medium	1	train

ClimateX: Dataset

Rows: 8093 statements

Source: AR6 WGI, AR6 WGII, AR6 WGIII

confidence: low, medium, high, very high

split: train(7794), test(300)

ClimateX: Github

https://github.com/rlacombe/ClimateX/blob/main/dsp_zeroshot_experiments.ipynb

- Model setting
- Loading the dataset
- Defining templates
- Defining the task
- Experiment
- Saving experiment results
- Precision, recall, and F1 score
- Over/under confidence assessment

ClimateX: Defining templates

Example:

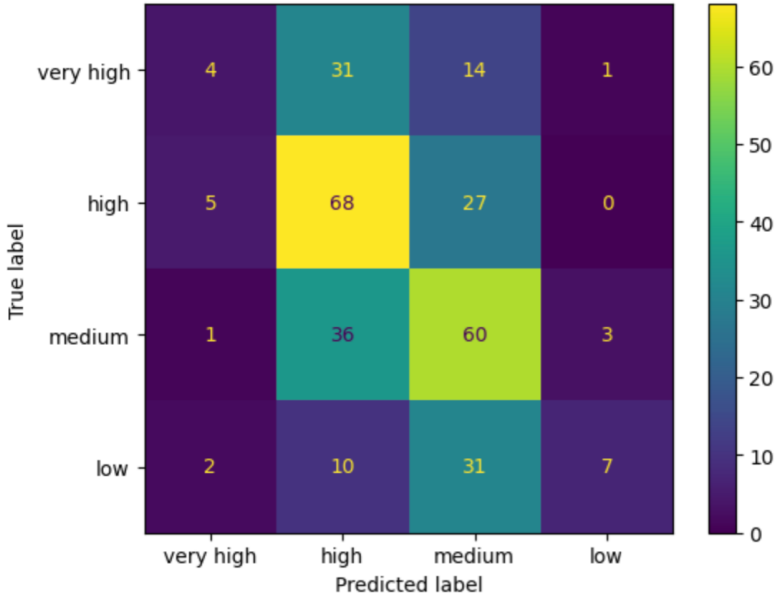
In [15]:

```
ex = dsp.Example(
    input=ipcc_train[0]['input'], label=ipcc_train[0]['label'])

ex.demos=dsp.sample(ipcc_train, 0)

print(zero_shot_template(ex))
```

```
You are a knowledgeable climate science assistant trained to assess the confidence level associated with various s
tatements about climate change.

You will be presented with a statement about climate science, climate impacts or climate change mitigation which i
s retrieved or paraphrased from the IPCC AR6 WGI, WGII or WGIII assessment reports. Climate scientists have evalua
ted that statement as low confidence, medium confidence, high confidence, or very high confidence, based on eviden
ce (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', 'very high'. If you don't know, you can r
espond 'I don't know'.

---

Follow the following format.

Statement: ${a short statement about climate.}
Confidence: ${must be *only*: 'low', 'medium', 'high', 'very high'}

---

Statement: Since 2011 (measurements reported in AR5), concentrations have continued to increase in the atmosphere,
reaching annual averages of 410 parts per million (ppm) for carbon dioxide (CO 2), 1866 parts per billion (ppb) fo
r methane (CH 4), and 332 ppb for nitrous oxide (N 2O) in 2019.6 Land and ocean have taken up a near-constant prop
ortion (globally about 56% per year) of CO 2 emissions from human activities over the past six decades, with regio
nal differences
Confidence:
```

CliamteX result: Dataframe

statement_i	report	page_num	sent_num	statement	confidence	score	split	prediction	correct
3	AR6_WGI	24	2	Since 1750, i	very high	3	test	very high	TRUE
42	AR6_WGI	37	16	Over the nex	low	0	test	high	FALSE
77	AR6_WGI	47	7	By the end o	high	2	test	high	TRUE
81	AR6_WGI	62	2	Over the pas	medium	1	test	high	FALSE
86	AR6_WGI	63	8	The paleo co	high	2	test	very high	FALSE
98	AR6_WGI	65	30	These higher	medium	1	test	medium	TRUE
151	AR6_WGI	85	31	Model estim	low	0	test	medium	FALSE
157	AR6_WGI	87	27	Projected cha	medium	1	test	medium	TRUE
162	AR6_WGI	90	4	A long-term	medium	1	test	very high	FALSE
165	AR6_WGI	90	14	Ocean warm	medium	1	test	very high	FALSE
190	AR6_WGI	93	13	The total Ant	very high	3	test	high	FALSE
197	AR6_WGI	93	34	Since AR5, th	high	2	test	very high	FALSE
233	AR6_WGI	101	6	Water cycle	high	2	test	high	TRUE
237	AR6_WGI	101	11	Global land p	medium	1	test	high	FALSE
282	AR6_WGI	116	13	The largest c	medium	1	test	high	FALSE
322	AR6_WGI	122	19	For global w	low	0	test	high	FALSE




```
Macro F1 score: 0.35772096872812476
Weighted F1 score: 0.41720454917319505
Accuracy (total): 0.4633333333333333
              precision    recall  f1-score   support

        high     0.4690    0.6800    0.5551       100
         low     0.6364    0.1400    0.2295        50
      medium     0.4545    0.6000    0.5172       100
   very high     0.3333    0.0800    0.1290        50

    accuracy                         0.4633       300
   macro avg     0.4733    0.3750    0.3577       300
weighted avg     0.4695    0.4633    0.4172       300

confidence
high         100
medium       100
very high     50
low           50
Name: count, dtype: int64
```

ClimateX: Precision, recall, and F1 score

Precision: colunm mean

Recall: row mean

F1 score: $\frac{2 \times Precision \times Recall}{Precision + Recall}$

ClimateX result

Details:
https://github.com/rlacombe/ClimateX/blob/main/results/iccs-zeroshot/gpt4-zeroshot-temp0-2023-06-09.csv

D Appendix: Classifier Results Table

Table 4 presents the precision, recall, and F1 score for each classifier, as well as support (number of sentences for which the model answered with a valid confidence label). Note that the 'very high' class and the 'low' class each have 50 total sentences, while the 'high' and 'medium' classes each have 100, for a total of 300 sentences in the test set.

Models		**GPT-3.5-turbo**		**GPT-4**		**Cohere Command XL**	
Setting		Zero-shot	Few-shot	Zero-shot	Few-shot	Zero-shot	Few-shot
'very high'	Precision	0.500	0.476	0.428	0.375	0.221	0.238
	Recall	0.146	0.208	0.120	0.180	0.300	0.592
	F1	0.226	0.290	0.188	0.243	0.254	0.339
	Support	48	48	50	50	50	49
'high'	Precision	0.504	0.485	0.472	0.475	0.332	0.383
	Recall	0.582	0.505	0.680	0.660	0.760	0.546
	F1	0.540	0.495	0.557	0.552	0.462	0.450
	Support	98	99	100	100	100	99
'medium'	Precision	0.389	0.389	0.410	0.466	0.500	0.0
	Recall	0.636	0.616	0.570	0.610	0.010	0.0
	F1	0.483	0.477	0.477	0.528	0.020	0.0
	Support	99	99	100	100	100	100
'low'	Precision	0.167	0.143	0.667	0.833	1.000	0.353
	Recall	0.020	0.041	0.040	0.100	0.020	0.245
	F1	0.036	0.064	0.076	0.179	0.039	0.289
	Support	50	49	50	50	50	49
Aggregate	**Accuracy**	**0.434**	**0.417**	**0.443**	**0.470**	**0.310**	**0.320**
	Macro F1	0.321	0.331	0.324	0.376	0.194	0.270
	Weighted F1	0.384	0.384	0.389	0.430	0.209	0.254
	Support	295	295	300	300	300	297

Table 4: Detailed results: Model classification performance results for the 3 models we assessed in both the zero shot and few shot setting. Reported metrics: accuracy, weighted and macro F1 score, and class-wise recall, precision, and F1 metrics.

Our ChatIAMs

Source:

1. Paper Draft: https://www.overleaf.com/1225397724nhhpgfjszczg#a93b66
2. Github Code: https://github.com/yl1127/Academic-projects/blob/main/yl_Climate_LLM/yl_Climate_0515.ipynb
3. Dataset: https://huggingface.co/datasets/rlacombe/ClimateX
4. IAMs(Integrated Assessment Modelling): https://www.ipcc.ch/report/ar6/wg3/downloads/report/IPCC_AR6_WGIII_Annex-III.pdf
5. https://openscm-runner.readthedocs.io/en/latest/notebooks/magicc/run-magicc.html

IAMs(Integrated Assessment Modelling) Result

Climate Variables(10)

- Surface Air Temperature Change,
- Atmospheric Concentrations|CO2,
- Effective Radiative Forcing,
- CO2,
- Aerosols,
- Direct Effect|BC,
- Direct Effect|OC,
- Direct Effect|SOx,
- Direct Effect and Indirect Effect
- Sea Level Change

- **Time**: 1950-2100
- **Scenarios**: ssp119, ssp126, ssp245, ssp370, ssp460 and ssp585
- **Confidence interval**

IAMs(Integrated Assessment Modelling) Result

model	quantile	region	scenario	unit	variable	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963
AIM/CGE	0.005	World	ssp370	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
AIM/CGE	0.005	World	ssp370-lowN	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
AIM/CGE	0.005	World	ssp370-lowN	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
GCAM4	0.005	World	ssp434	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
GCAM4	0.005	World	ssp460	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
IMAGE	0.005	World	ssp119	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
IMAGE	0.005	World	ssp126	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
MESSAGE-G	0.005	World	ssp245	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
REMIND-MA	0.005	World	ssp534-over	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
REMIND-MA	0.005	World	ssp585	ppm	Atmospheric	312.821	313.014	313.342	313.73	314.095	314.415	314.698	314.992	315.345	315.807	316.625	317.299	318.044	318.65
AIM/CGE	0.005	World	ssp370	W / m^2	Effective Rad	0.80966205	0.81077522	0.79057856	0.79139343	0.79597298	0.81300747	0.82454413	0.84433543	0.90661575	0.89860682	0.89155208	0.73754744	0.45263293	0.15029892
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	0.79121384	0.79394895	0.77288623	0.77322946	0.77723483	0.79357729	0.80299683	0.82074545	0.88454495	0.87629651	0.86844606	0.71061633	0.42641029	0.12150809
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	0.79121482	0.79395029	0.77288766	0.7732309	0.77723634	0.79357887	0.80299848	0.82074713	0.88454667	0.8762981	0.86844763	0.71061809	0.42641207	0.1215099
GCAM4	0.005	World	ssp434	W / m^2	Effective Rad	0.78568667	0.78775469	0.76663331	0.7669466	0.77094004	0.78668819	0.7956232	0.8131947	0.87686123	0.86835434	0.86020667	0.70175987	0.4171346	0.11205342
GCAM4	0.005	World	ssp460	W / m^2	Effective Rad	0.78568667	0.78775469	0.76663331	0.7669466	0.77094004	0.78668819	0.7956232	0.8131947	0.87686123	0.86835434	0.86020667	0.70175987	0.4171346	0.11205342
IMAGE	0.005	World	ssp119	W / m^2	Effective Rad	0.75608834	0.76328763	0.74011381	0.73918868	0.7418622	0.75706093	0.76218838	0.77510625	0.84243892	0.83392309	0.82469352	0.65677924	0.37498351	0.06424094
IMAGE	0.005	World	ssp126	W / m^2	Effective Rad	0.75735758	0.764652	0.74147711	0.74055438	0.74323359	0.75854553	0.76377905	0.77673264	0.84408932	0.83564646	0.82649084	0.65864162	0.37692131	0.06621139
MESSAGE-G	0.005	World	ssp245	W / m^2	Effective Rad	0.7671681	0.77170857	0.74951881	0.74915014	0.75250786	0.76755881	0.77421458	0.7892204	0.85475033	0.84597021	0.8370461	0.67307447	0.38965827	0.08147979
REMIND-MA	0.005	World	ssp534-over	W / m^2	Effective Rad	0.75592244	0.76001522	0.73784248	0.73752714	0.74067671	0.75516356	0.76091636	0.77564355	0.84113219	0.8318951	0.82236415	0.65891232	0.37515457	0.0666921
REMIND-MA	0.005	World	ssp585	W / m^2	Effective Rad	0.75594287	0.76003582	0.73786316	0.73754789	0.74069753	0.75518448	0.76093738	0.77566469	0.8411535	0.8319167	0.82238608	0.6589345	0.37517677	0.06671422
AIM/CGE	0.005	World	ssp370	W / m^2	Effective Rad	-0.5304396	-0.5001678	-0.5202375	-0.5292447	-0.5436614	-0.5637574	-0.6206919	-0.6680614	-0.6300215	-0.635823	-0.6586038	-0.7385091	-0.7297062	-0.7928036
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	-0.5551753	-0.5234276	-0.5444155	-0.5538431	-0.568907	-0.5899676	-0.649566	-0.6991544	-0.6593253	-0.665441	-0.6893015	-0.7730214	-0.7638204	-0.8298716
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	-0.5551752	-0.5234275	-0.5444154	-0.553843	-0.5689069	-0.5899674	-0.6495658	-0.6991543	-0.6593251	-0.6654409	-0.6893014	-0.7730213	-0.7638203	-0.8298715
GCAM4	0.005	World	ssp434	W / m^2	Effective Rad	-0.556515	-0.5251568	-0.5461263	-0.5556163	-0.5706129	-0.592069	-0.6517806	-0.7014343	-0.661906	-0.6683134	-0.6923157	-0.7765654	-0.7676006	-0.8335888
GCAM4	0.005	World	ssp460	W / m^2	Effective Rad	-0.556515	-0.5251568	-0.5461263	-0.5556163	-0.5706129	-0.592069	-0.6517806	-0.7014343	-0.661906	-0.6683134	-0.6923157	-0.7765654	-0.7676006	-0.8335888
IMAGE	0.005	World	ssp119	W / m^2	Effective Rad	-0.6176804	-0.5820325	-0.6053387	-0.6157834	-0.6324793	-0.655887	-0.7222108	-0.7773758	-0.7327347	-0.7394581	-0.7659657	-0.8596647	-0.8493915	-0.9229873
IMAGE	0.005	World	ssp126	W / m^2	Effective Rad	-0.6172275	-0.5815023	-0.6048156	-0.6152491	-0.6319481	-0.6552797	-0.7215722	-0.7767227	-0.7320153	-0.7386745	-0.7651461	-0.8587762	-0.8484662	-0.9220806
MESSAGE-G	0.005	World	ssp245	W / m^2	Effective Rad	-0.5899441	-0.5565909	-0.5787846	-0.5888308	-0.6046579	-0.6274759	-0.690784	-0.7434131	-0.7013667	-0.708162	-0.7335968	-0.8234114	-0.8139557	-0.8839538
REMIND-MA	0.005	World	ssp534-over	W / m^2	Effective Rad	-0.59386	-0.5608477	-0.5829814	-0.5930545	-0.6090188	-0.6320715	-0.695606	-0.748323	-0.7067081	-0.7138575	-0.7395438	-0.8289024	-0.8194307	-0.8892413
REMIND-MA	0.005	World	ssp585	W / m^2	Effective Rad	-0.59386	-0.5608477	-0.5829814	-0.5930545	-0.6090189	-0.6320715	-0.695606	-0.748323	-0.7067081	-0.7138575	-0.7395438	-0.8289024	-0.8194307	-0.8892413
AIM/CGE	0.005	World	ssp370	W / m^2	Effective Rad	-0.1057173	-0.111232	-0.1127764	-0.1148976	-0.1162407	-0.1259323	-0.1344366	-0.1396555	-0.1395005	-0.143876	-0.1484634	-0.1698231	-0.1735158	-0.1789923
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	-0.1106161	-0.1163264	-0.1179237	-0.1201438	-0.1215233	-0.1316937	-0.1406	-0.1460691	-0.1458942	-0.1505174	-0.1553348	-0.1777748	-0.1816529	-0.187392
AIM/CGE	0.005	World	ssp370-lowN	W / m^2	Effective Rad	-0.110616	-0.1163263	-0.1179236	-0.1201437	-0.1215232	-0.1316936	-0.1405999	-0.1460689	-0.1458941	-0.1505173	-0.1553347	-0.1777746	-0.1816527	-0.1873919
GCAM4	0.005	World	ssp434	W / m^2	Effective Rad	-0.1139736	-0.1199034	-0.1215703	-0.1238855	-0.1252599	-0.1358751	-0.1451247	-0.1508593	-0.1508054	-0.155727	-0.1607726	-0.1840205	-0.1880755	-0.1940253

Our method

In [318…

```
ex = Example(
    input=ipcc_train[0]['input'], table = function_response, label=ipcc_train[0]['label'])

ex.demos=sample(ipcc_train, 0)

print(zero_shot_template(ex))
```

```
You are a knowledgeable climate science assistant trained to assess the confidence level associated with various s
tatements about climate change.

You will be presented with a statement about climate science, climate impacts or climate change mitigation which i
s retrieved or paraphrased from the IPCC AR6 WGI, WGII or WGIII assessment reports. Climate scientists have evalua
ted that statement as low confidence, medium confidence, high confidence, or very high confidence, based on eviden
ce (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', 'very high'. If you don't know, you can r
espond 'I don't know'.

---

Follow the following format.

Statement: ${a short statement about climate.}
IAMs output: ${a json table about dataset related with statment above.}
Confidence: ${must be *only*: 'low', 'medium', 'high', 'very high'}

---

Statement: Since 2011 (measurements reported in AR5), concentrations have continued to increase in the atmosphere,
reaching annual averages of 410 parts per million (ppm) for carbon dioxide (CO 2), 1866 parts per billion (ppb) fo
r methane (CH 4), and 332 ppb for nitrous oxide (N 2O) in 2019.6 Land and ocean have taken up a near-constant prop
ortion (globally about 56% per year) of CO 2 emissions from human activities over the past six decades, with regio
nal differences
IAMs output: {"model":{"19":"REMIND-MAGPIE","119":"REMIND-MAGPIE","219":"REMIND-MAGPIE","319":"REMIND-MAGPIE","41
9":"REMIND-MAGPIE"},"quantile":{"19":0.005,"119":0.025,"219":0.5,"319":0.975,"419":0.995},"variable":{"19":"Effect
ive Radiative Forcing","119":"Effective Radiative Forcing","219":"Effective Radiative Forcing","319":"Effective Ra
diative Forcing","419":"Effective Radiative Forcing"},"unit":{"19":"W \/ m^2","119":"W \/ m^2","219":"W \/ m^2","3
19":"W \/ m^2","419":"W \/ m^2"},"scenario":{"19":"ssp585","119":"ssp585","219":"ssp585","319":"ssp585","419":"ssp
585"},"2100":{"19":9.231232444,"119":9.237379018,"219":9.3321387,"319":9.427793228,"419":9.433997846}}
Confidence:
```

Our result

statement_i	report	page_num	sent_num	statement	confidence	score	split	tables	prediction	correct
3	AR6_WGI	24	2	Since 1750, i	very high	3	test		very high	TRUE
42	AR6_WGI	37	16	Over the nex	low	0	test	{"model":{"5	very high	FALSE
77	AR6_WGI	47	7	By the end o	high	2	test	{"model":{"5	very high	FALSE
81	AR6_WGI	62	2	Over the pas	medium	1	test		high	FALSE
86	AR6_WGI	63	8	The paleo co	high	2	test		high	TRUE
98	AR6_WGI	65	30	These higher	medium	1	test		medium	TRUE
151	AR6_WGI	85	31	Model estim	low	0	test		medium	FALSE
157	AR6_WGI	87	27	Projected cha	medium	1	test		medium	TRUE
162	AR6_WGI	90	4	A long-term	medium	1	test	{"model":{"9	high	FALSE
165	AR6_WGI	90	14	Ocean warm	medium	1	test	{"model":{"9	high	FALSE
190	AR6_WGI	93	13	The total Ant	very high	3	test	{"model":{"5	high	FALSE
197	AR6_WGI	93	34	Since AR5, th	high	2	test	{"model":{"5	high	TRUE
233	AR6_WGI	101	6	Water cycle	high	2	test	{"model":{"9	high	TRUE
237	AR6_WGI	101	11	Global land p	medium	1	test	{"model":{"9	high	FALSE
282	AR6_WGI	116	13	The largest c	medium	1	test	{"model":{"9	medium	TRUE
322	AR6_WGI	122	19	For global w	low	0	test	{"model":{"9	medium	FALSE
361	AR6_WGI	134	16	At global and	medium	1	test	{"model":{"9	medium	TRUE

Our result

Details: https://github.com/yl1127/Academic-projects/blob/main/yl_Climate_LLM/ChatIAMs/

Our Result:

level2	**GPT-3.5-turbo**	**GPT-4**	**GPT-4-turbo**	**GPT-3.5-turbo**	**GPT-4**	**GPT-4-turbo**	**GPT-4-turbo**
Settings	Zero-shot	Zero-shot	Zero-shot	few-shots	few-shots	few-shots(random)	few-shots(semantic search)
Accuracy	41.5	45.2	46.3	37.9	43.3	47.7	46.3
Note				idk:102			

ClimateX:

ClimateX	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Settings	Zero-shot	Zero-shot	few-shots	few-shots
Accuracy	43.4	44.3	41.7	47.0

Compare with ClimateX

Initialize Reactive Jupyter | Sync all Stale code

```
samples_compare['pred==pred_gpt4'].value_counts()
```

[12] ✓ 0.0s

```
pred==pred_gpt4
True     222
False     78
Name: count, dtype: int64
```

30 wrong -> correct
24 correct -> wrong
24 wrong -> wrong

Few-shot

```
You are a knowledgeable climate science assistant trained to assess the confidence level associated with various statements a

You will be presented with a statement about climate science, climate impacts or climate change mitigation which is retrieved

Respond *only* with one of the following words: 'low', 'medium', 'high', 'very high'. If you don't know, you can respond 'I d

---

Follow the following format.

Statement: ${a short statement about climate.}
IAMs output: ${a json table from IAM climate modeling.}
Confidence: ${must be *only*: 'low', 'medium', 'high', 'very high'}

---

Statement: For example, pathways that lead to poverty reduction can have synergies with food security, water, gender, terrest
Confidence: very high

Statement: Increases in frequency, intensity and severity of droughts, floods and heatwaves, and continued sea level rise wil
Confidence: medium

Statement: Relative to 1995–2014, the likely global mean sea level rise by 2100 is 0.28–0.55 m under the very low GHG emissio
Confidence: low

Statement: By the end of the century, scenarios with very low and low GHG emissions would strongly limit the change of severa
IAMs output: {"model":{"510":"uSEM"},"quantile":{"510":0.5},"variable":{"510":"Sea Level Change"},"unit":{"510":"mm"},"scenar
Confidence:
```

Few-shots results

Our Result:

level2	**GPT-3.5-turbo**	**GPT-4**	**GPT-4-turbo**	**GPT-3.5-turbo**	**GPT-4**	**GPT-4-turbo**	**GPT-4-turbo**
Settings	Zero-shot	Zero-shot	Zero-shot	few-shots	few-shots	few-shots(random)	few-shots(semantic search)
Accuracy	41.5	45.2	46.3	37.9	43.3	47.7	46.3
Note				idk:102			

ClimateX:

ClimateX	**GPT-3.5-turbo**	**GPT-4**	**GPT-3.5-turbo**	**GPT-4**
Settings	Zero-shot	Zero-shot	few-shots	few-shots
Accuracy	43.4	44.3	41.7	47.0