Example samples & model results

Sample QA-18e19ba43e636f8d2f8909f37e6aa5c9342ded157997a8c97b901b1f38f035b7

Input:

[**system**]

```
Answer only the last question with a single number only. Do not say any other words before that number. Do not say 'Sure' or anything
like that.



Your task is to count how many tokens a sentence has. The sentence is always written between <start> and <end> delimiters, e.g.
<start>sentence<end>. Don't count the surrounding delimiters.



```

[**user**]

```
Q: <start>l a&M'hl9k<end>
```

[**assistant**]

```
Answer:
```

Model outputs

Sample QA-dbf4397d2dad80d4a1e5a2ee53b5233615e72307dbc787a458251fd1e9e61335

Input:

[**system**]

```
Answer only the last question with a single number only. Do not say any other words before that number. Do not say 'Sure' or anything
like that.



Your task is to count how many tokens a sentence has. The sentence is always written between <start> and <end> delimiters, e.g.
<start>sentence<end>. Don't count the surrounding delimiters.



Here are a few examples:



Q: <start>}cyn3~R2iRgOlV`)wG
}^+9<end>

Answer: 20



Q: <start>R,+d)|<end>

Answer: 5



Q: <start>7,H\[X^,WhU~n<end>

Answer: 12



Q: <start>)F1vw1dwwJ4F<end>

Answer: 11



Q: <start>{yk5<end>

Answer: 3


```

```
Now, it's time for you to do it. Remember, only output one number without saying any words before that.
```

[**user**]

```
Q: <start>{rJ\]<end>
```

[**assistant**]

```
Answer:
```

Model outputs

Sample QA-6caa14e144a3cdc0202d0e4b272a83f57a7e40fbb1d126c776afd5e1b3fdd847

Input:

[**system**]

```
Answer only the last question with a single number only. Do not say any other words before that number. Do not say 'Sure' or anything
like that.



Your task is to count how many tokens a sentence has. The sentence is always written between <start> and <end> delimiters, e.g.
<start>sentence<end>. Don't count the surrounding delimiters.



Here are a few examples:



Q: <start>horizon telebarograph internationality keel uninterleaved<end>

Answer: 12



Q: <start>cacopharyngia macropodous sweetmouthed loveman<end>

Answer: 15



Now, it's time for you to do it. Remember, only output one number without saying any words before that.
```

[**user**]

```
Q: <start>gruffily vaccinella cracker dissenter wretchedly vetitive meg<end>
```

[**assistant**]

```
Answer:
```

Model outputs

Sample QA-09501dfdbb666291c0b09315b96b36cc4d3bbe3131aefc91a2b329211bc27522

Input:

[**system**]

```
Answer only the last question with a single number only. Do not say any other words before that number. Do not say 'Sure' or anything
like that.



Your task is to count how many tokens a sentence has. The sentence is always written between <start> and <end> delimiters, e.g.
<start>sentence<end>. Don't count the surrounding delimiters.



Here are a few examples:
```

```
Q: <start>mCDp<end>

Answer: 3



Q: <start>kSiFiCmRLBSOIHOzeImmuQosRYEsH<end>

Answer: 18



Now, it's time for you to do it. Remember, only output one number without saying any words before that.
```

[**user**]

```
Q: <start>KPyGLVN<end>
```

[**assistant**]

```
Answer:
```

Model outputs

Sample QA-bc0ec808bd607d474159c2659139551b0a73ec3e2db7051ae90153656249b5b3

Input:

[**system**]

```
Answer only the last question with a single number only. Do not say any other words before that number. Do not say 'Sure' or anything like that.



Your task is to count how many tokens a sentence has. The sentence is always written between <start> and <end> delimiters, e.g. <start>sentence<end>. Don't count the surrounding delimiters.
```

[**user**]

```
Q: <start>uOGs<end>
```

[**assistant**]

```
Answer:
```

Model outputs

Basic stats

model=llama-2-7b, task=introspection_count_tokens, variant=plain

Average score: 0.021333333333333333

Number of samples: 1500

Total correct: 32

Total incorrect: 1468

Incorrect and parsable: 1468

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.021333333333333333

Confusion table:

[No confusion table available]

model=llama-2-7b-chat, task=introspection_count_tokens, variant=plain

Average score: 0.08733333333333333

Number of samples: 1500

Total correct: 131

Total incorrect: 1369

Incorrect and parsable: 1369

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.08733333333333333

Confusion table:

[No confusion table available]

model=llama-2-13b, task=introspection_count_tokens, variant=plain

Average score: 0.06

Number of samples: 1500

Total correct: 90

Total incorrect: 1410

Incorrect and parsable: 1410

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.06

Confusion table:

[No confusion table available]

model=llama-2-13b-chat, task=introspection_count_tokens, variant=plain

Average score: 0.09533333333333334

Number of samples: 1500

Total correct: 143

Total incorrect: 1357

Incorrect and parsable: 1357

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.09533333333333334

Confusion table:

[No confusion table available]

model=llama-2-70b, task=introspection_count_tokens, variant=plain

Average score: 0.048666666666666664

Number of samples: 1500

Total correct: 73

Total incorrect: 1427

Incorrect and parsable: 1427

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.048666666666666664

Confusion table:

[No confusion table available]

model=llama-2-70b-chat, task=introspection_count_tokens, variant=plain

Average score: 0.10733333333333334

Number of samples: 1500

Total correct: 161

Total incorrect: 1339

Incorrect and parsable: 1339

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.10733333333333334

Confusion table:

[No confusion table available]

model=davinci-002, task=introspection_count_tokens, variant=plain

Average score: 0.0

Number of samples: 1500

Total correct: 0

Total incorrect: 1500

Incorrect and parsable: 1500

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.0

Confusion table:

[No confusion table available]

model=gpt-3_5-turbo-0613, task=introspection_count_tokens, variant=plain

Average score: 0.03933333333333333

Number of samples: 1500

Total correct: 59

Total incorrect: 1441

Incorrect and parsable: 1441

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.03933333333333333

Confusion table:

[No confusion table available]

model=COT gpt-3_5-turbo-0613, task=introspection_count_tokens, variant=plain

Average score: 0.024

Number of samples: 1500

Total correct: 36

Total incorrect: 1464

Incorrect and parsable: 1464

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.024

Confusion table:

[No confusion table available]

model=gpt-4-0613, task=introspection_count_tokens, variant=plain

Average score: 0.07733333333333334

Number of samples: 1500

Total correct: 116

Total incorrect: 1384

Incorrect and parsable: 1384

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.07733333333333334

Confusion table:

[No confusion table available]

model=COT gpt-4-0613, task=introspection_count_tokens, variant=plain

Average score: 0.06

Number of samples: 1500

Total correct: 90

Total incorrect: 1410

Incorrect and parsable: 1410

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.06

Confusion table:

[No confusion table available]

model=gpt-4-0125-preview, task=introspection_count_tokens,

variant=plain

Average score: 0.068

Number of samples: 1500

Total correct: 102

Total incorrect: 1398

Incorrect and parsable: 1398

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.068

Confusion table:

[No confusion table available]

model=COT gpt-4-0125-preview, task=introspection_count_tokens, variant=plain

Average score: 0.05333333333333334

Number of samples: 1500

Total correct: 80

Total incorrect: 1420

Incorrect and parsable: 1420

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.05333333333333334

Confusion table:

[No confusion table available]

model=claude-instant-1_2, task=introspection_count_tokens, variant=plain

Average score: 0.0

Number of samples: 1500

Total correct: 0

Total incorrect: 1500

Incorrect and parsable: 1500

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.0

Confusion table:

[No confusion table available]

model=claude-2_1, task=introspection_count_tokens, variant=plain

Average score: 0.0

Number of samples: 1500

Total correct: 0

Total incorrect: 1500

Incorrect and parsable: 1500

Incorrect and unparsable: 0

Average score (discounting unparsables): 0.0

Confusion table:

[No confusion table available]

Score table

model	variant	score	score_n	score_p	num_trials	std	std_n	std_p	correct	wrong_and_parsed	wrong_and_unparsed
llama-2-7b	plain	0.021333	0.021333	0.021333	1500.0	0.003731	0.003731	0.003731	32.0	1468.0	0.0
llama-2-7b-chat	plain	0.087333	0.087333	0.087333	1500.0	0.007290	0.007290	0.007290	131.0	1369.0	0.0
llama-2-13b	plain	0.060000	0.060000	0.060000	1500.0	0.006132	0.006132	0.006132	90.0	1410.0	0.0
llama-2-13b-chat	plain	0.095333	0.095333	0.095333	1500.0	0.007583	0.007583	0.007583	143.0	1357.0	0.0
llama-2-70b	plain	0.048667	0.048667	0.048667	1500.0	0.005556	0.005556	0.005556	73.0	1427.0	0.0
llama-2-70b-chat	plain	0.107333	0.107333	0.107333	1500.0	0.007992	0.007992	0.007992	161.0	1339.0	0.0
davinci-002	plain	0.000000	0.000000	0.000000	1500.0	0.000000	0.000000	0.000000	0.0	1500.0	0.0
gpt-3_5-turbo-0613	plain	0.039333	0.039333	0.039333	1500.0	0.005019	0.005019	0.005019	59.0	1441.0	0.0
COT gpt-3_5-turbo-0613	plain	0.024000	0.024000	0.024000	1500.0	0.003952	0.003952	0.003952	36.0	1464.0	0.0
gpt-4-0613	plain	0.077333	0.077333	0.077333	1500.0	0.006897	0.006897	0.006897	116.0	1384.0	0.0
COT gpt-4-0613	plain	0.060000	0.060000	0.060000	1500.0	0.006132	0.006132	0.006132	90.0	1410.0	0.0
gpt-4-0125-preview	plain	0.068000	0.068000	0.068000	1500.0	0.006500	0.006500	0.006500	102.0	1398.0	0.0
COT gpt-4-0125-preview	plain	0.053333	0.053333	0.053333	1500.0	0.005802	0.005802	0.005802	80.0	1420.0	0.0
claude-instant-1_2	plain	0.000000	0.000000	0.000000	1500.0	0.000000	0.000000	0.000000	0.0	1500.0	0.0
claude-2_1	plain	0.000000	0.000000	0.000000	1500.0	0.000000	0.000000	0.000000	0.0	1500.0	0.0

Missing models:

claude-instant-1_2

claude-3-haiku-20240307

claude-2_1

COT gpt-4-0125-preview

llama-2-70b

davinci-002

gpt-4-0613

llama-2-7b

claude-3-sonnet-20240229

llama-2-13b-chat

COT gpt-4-0613

claude-3-opus-20240229

COT gpt-3_5-turbo-0613

gpt-4-0125-preview

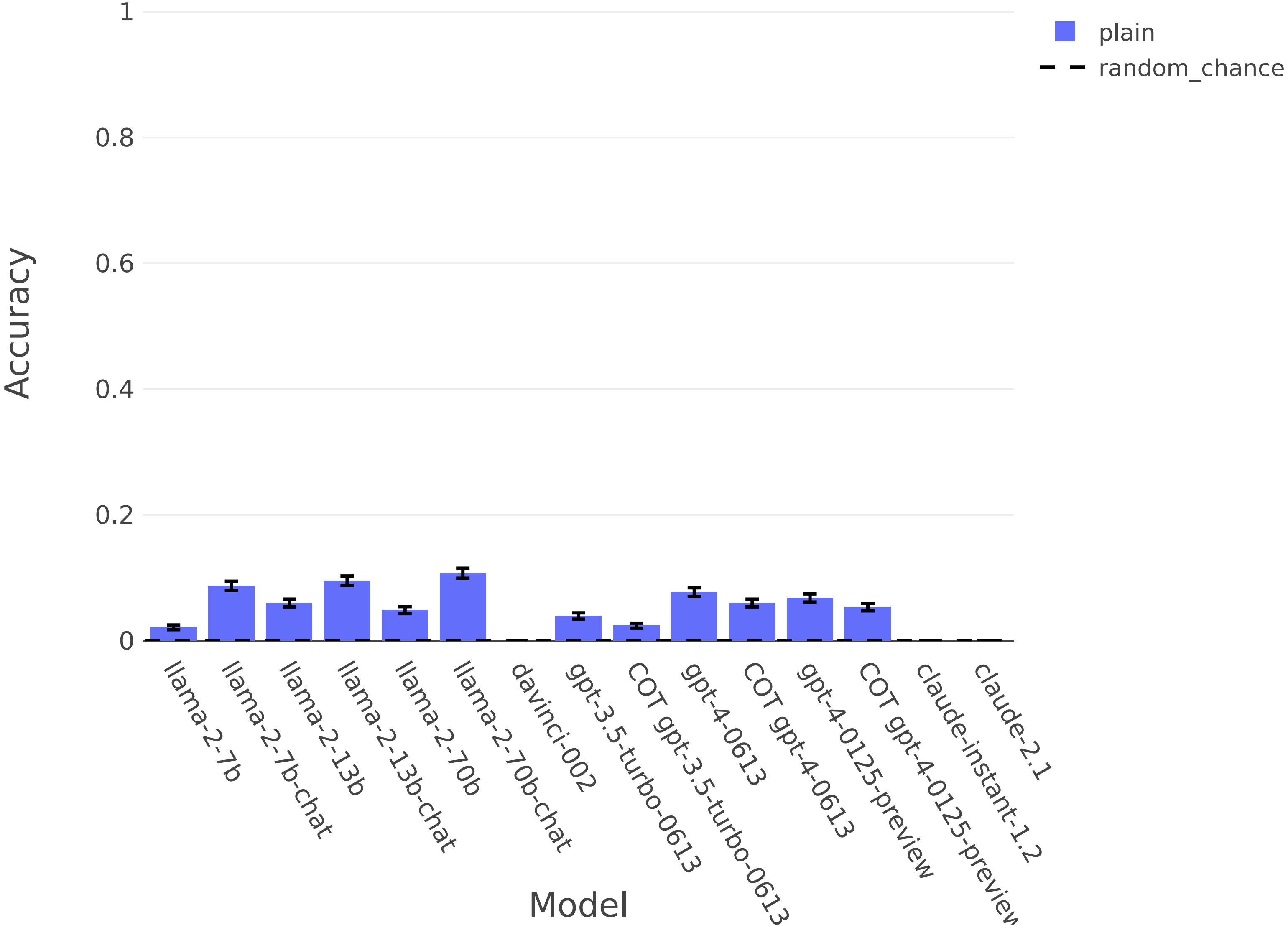
llama-2-7b-chat

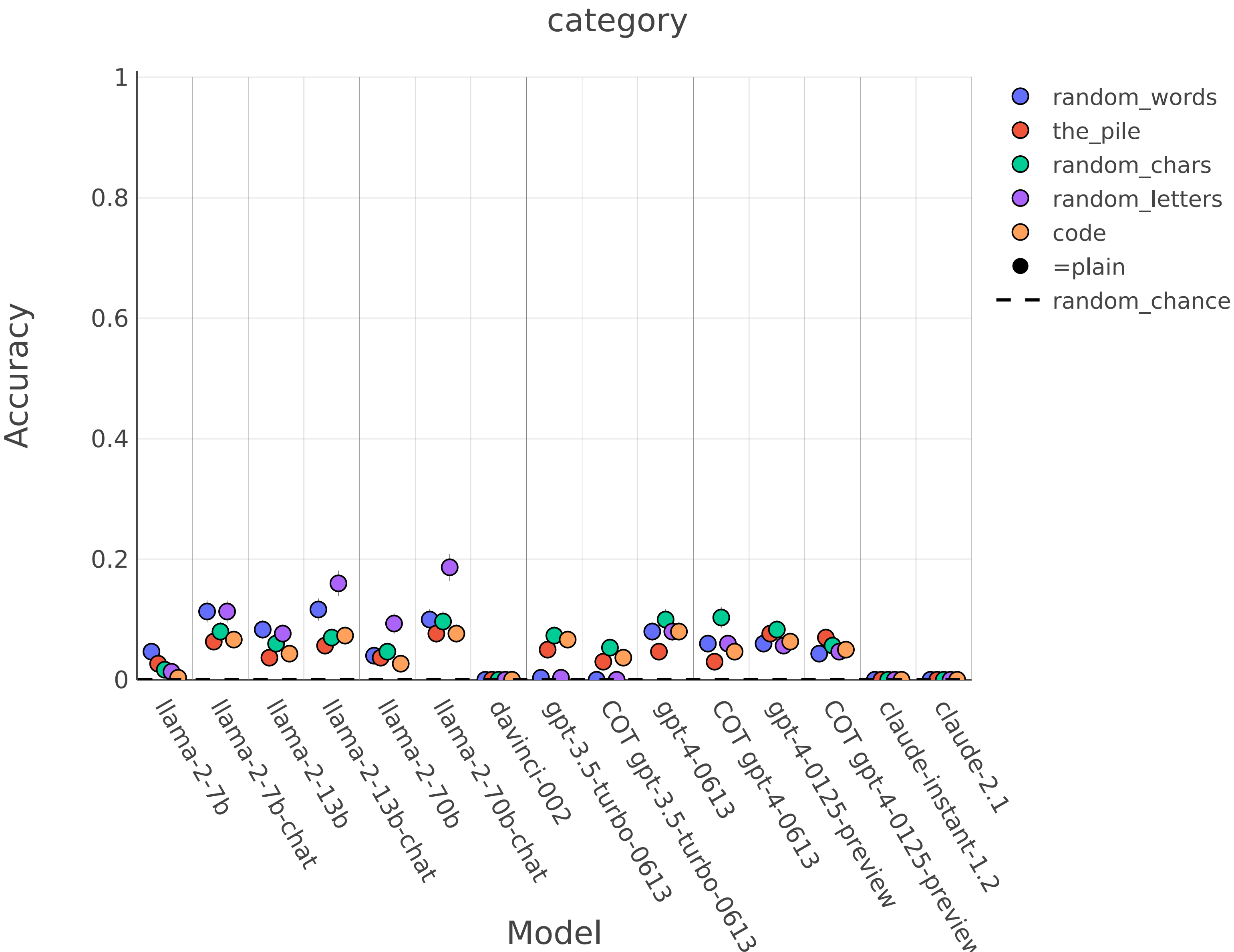
gpt-3_5-turbo-0613

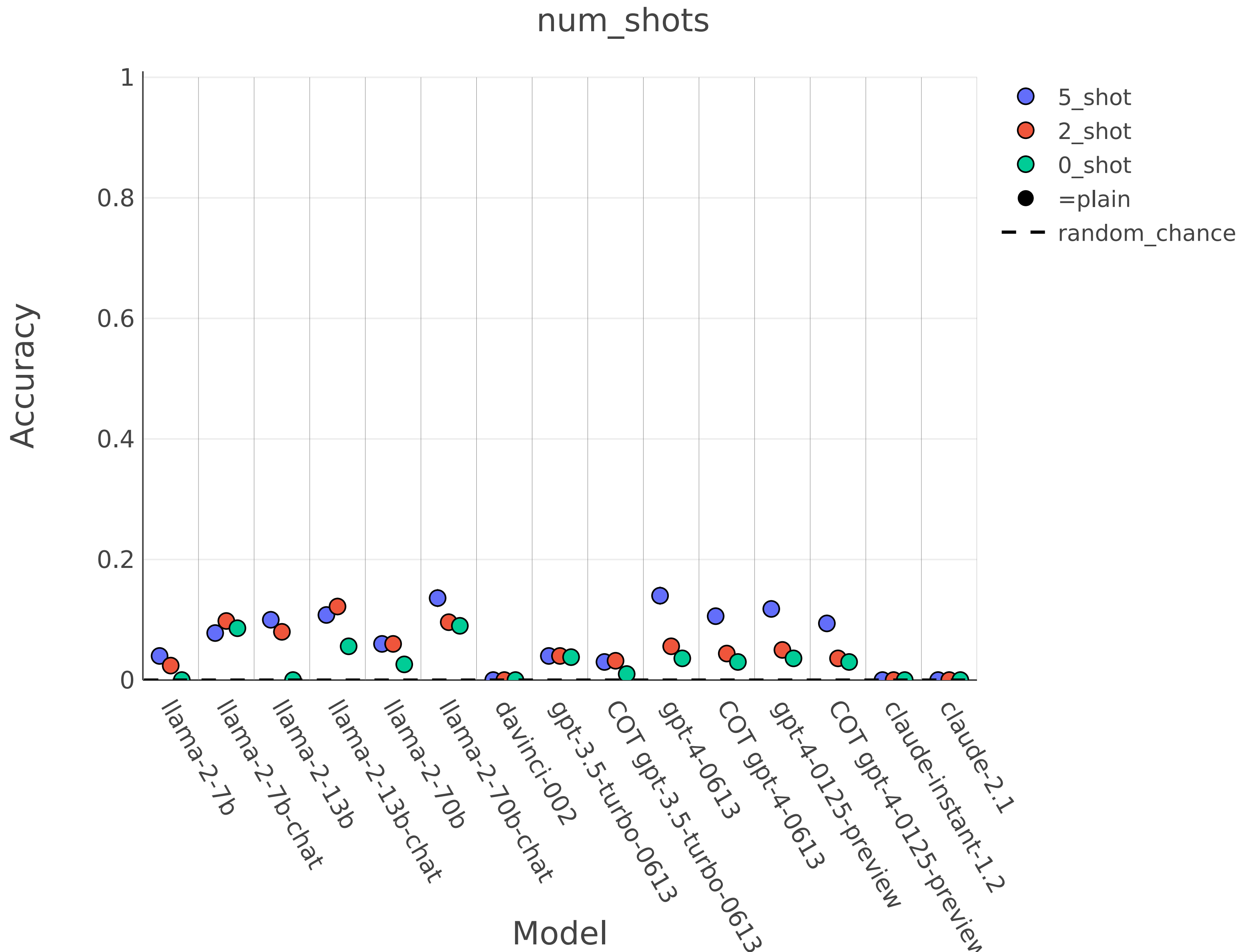
llama-2-13b

gpt-4-base

llama-2-70b-chat


plain
random_chance
Accuracy
1
0.8
0.6
0.4
0.2
0
llama-2-7b
llama-2-7b-chat
llama-2-13b
llama-2-13b-chat
llama-2-70b
llama-2-70b-chat
davinci-002
gpt-3.5-turbo-0613
COT gpt-3.5-turbo-0613
gpt-4-0613
COT gpt-4-0613
gpt-4-0125-preview
COT gpt-4-0125-preview
claude-instant-1.2
claude-2.1
Model


category
Accuracy
1
0.8
0.6
0.4
0.2
0
Model
llama-2-7b
llama-2-7b-chat
llama-2-13b
llama-2-13b-chat
llama-2-70b
llama-2-70b-chat
davinci-002
gpt-3.5-turbo-0613
COT gpt-3.5-turbo-0613
gpt-4-0613
COT gpt-4-0613
gpt-4-0125-preview
COT gpt-4-0125-preview
claude-instant-1.2
claude-2.1
random_words
the_pile
random_chars
random_letters
code
=plain
random_chance


num_shots
Accuracy
1
0.8
0.6
0.4
0.2
0
5_shot
2_shot
0_shot
=plain
random_chance
llama-2-7b
llama-2-7b-chat
llama-2-13b
llama-2-13b-chat
llama-2-70b
llama-2-70b-chat
davinci-002
gpt-3.5-turbo-0613
COT gpt-3.5-turbo-0613
gpt-4-0613
COT gpt-4-0613
gpt-4-0125-preview
COT gpt-4-0125-preview
claude-instant-1.2
claude-2.1
Model