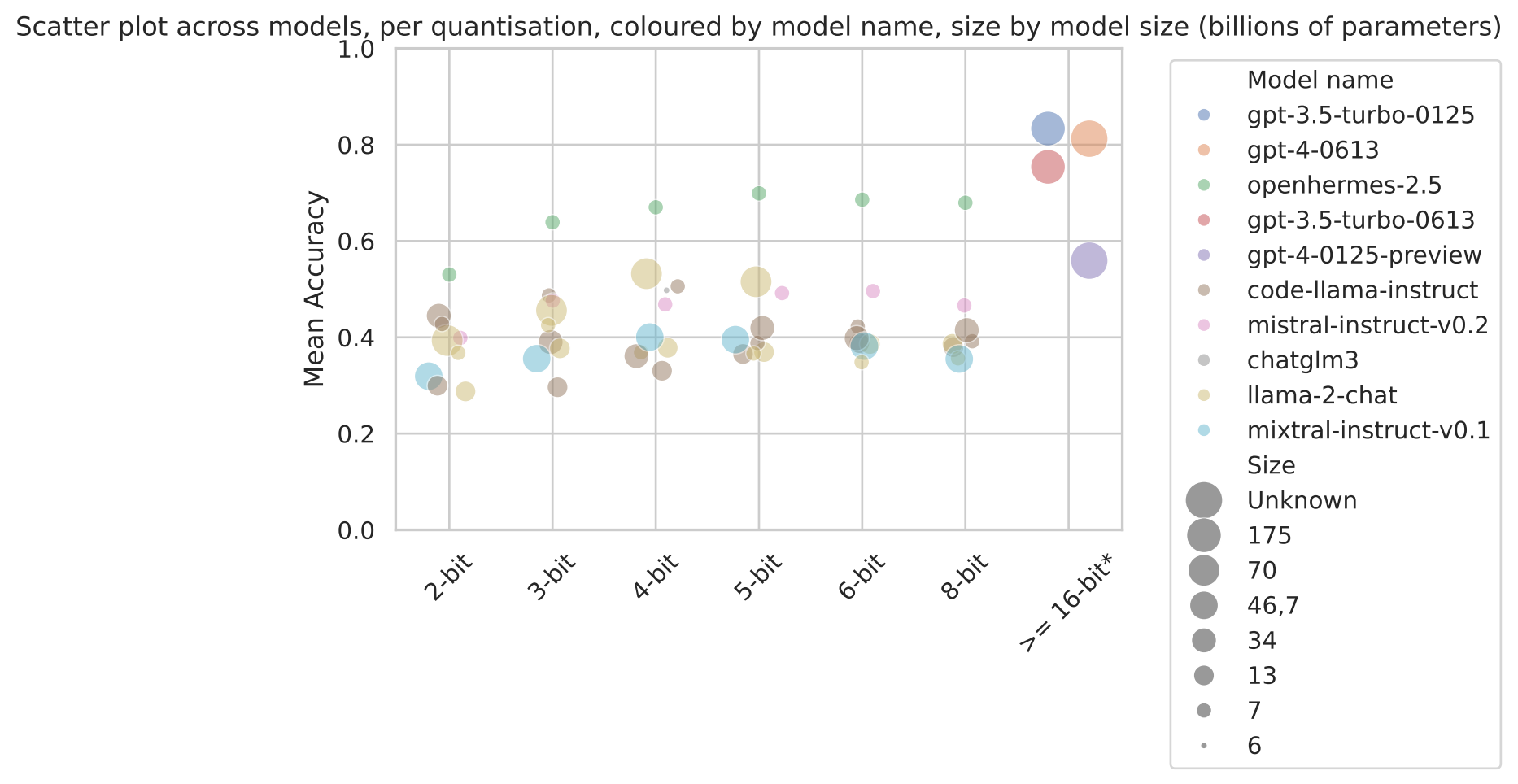
Scatter plot across models, per quantisation, coloured by model name, size by model size (billions of parameters)
Mean Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
2-bit
3-bit
4-bit
5-bit
6-bit
8-bit
>= 16-bit*
Model name
gpt-3.5-turbo-0125
gpt-4-0613
openhermes-2.5
gpt-3.5-turbo-0613
gpt-4-0125-preview
code-llama-instruct
mistral-instruct-v0.2
chatglm3
llama-2-chat
mixtral-instruct-v0.1
Size
Unknown
175
70
46,7
34
13
7
6