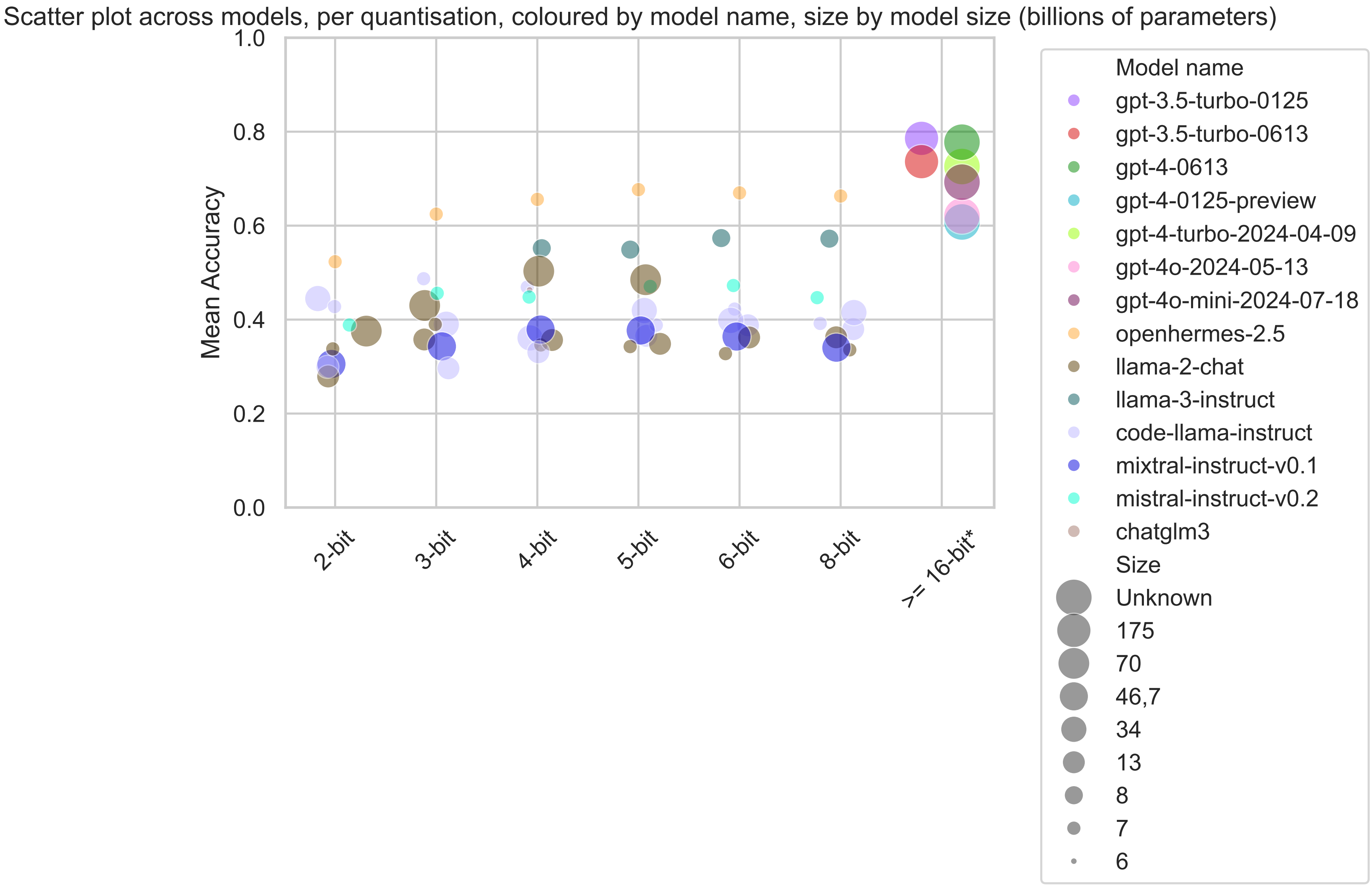
Scatter plot across models, per quantisation, coloured by model name, size by model size (billions of parameters)
Mean Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
2-bit
3-bit
4-bit
5-bit
6-bit
8-bit
>= 16-bit*
Model name
gpt-3.5-turbo-0125
gpt-3.5-turbo-0613
gpt-4-0613
gpt-4-0125-preview
gpt-4-turbo-2024-04-09
gpt-4o-2024-05-13
gpt-4o-mini-2024-07-18
openhermes-2.5
llama-2-chat
llama-3-instruct
code-llama-instruct
mixtral-instruct-v0.1
mistral-instruct-v0.2
chatglm3
Size
Unknown
175
70
46,7
34
13
8
7
6