Scatter plot across models, per quantisation, coloured by model name, size by model size (billions of parameters) Model name gpt-3.5-turbo-0125 8.0 gpt-3.5-turbo-0613 apt-4-0613 Mean Accuracy gpt-4-0125-preview 0.6 gpt-4o-2024-05-13 0.4 openhermes-2.5 llama-2-chat llama-3-instruct 0.2 code-llama-instruct mixtral-instruct-v0.1 mistral-instruct-v0.2 0.0 8.Dit chatglm3 A'O'A 3.bit Sibit 6.bit

