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-4-turbo-2024-04-09 gpt-4o-2024-05-13 gpt-4o-mini-2024-07-18 0.4 openhermes-2.5 llama-2-chat Ilama-3-instruct 0.2 code-llama-instruct mixtral-instruct-v0.1 mistral-instruct-v0.2 0.0 chatglm3 2,51 AJOIL 3.bit 5.bit 6.bit 8.bit Size Unknown 175 70 46,7 34 13 8

6