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 openhermes-2.5 Mean Accuracy gpt-4-0125-preview 0.6 gpt-4-0613 gpt-4o-2024-05-13 llama-2-chat 0.4 llama-3-instruct code-llama-instruct mixtral-instruct-v0.1 0.2 mistral-instruct-v0.2 chatglm3 0.0 Size Unknown 175 70 46,7 34 13

8

6