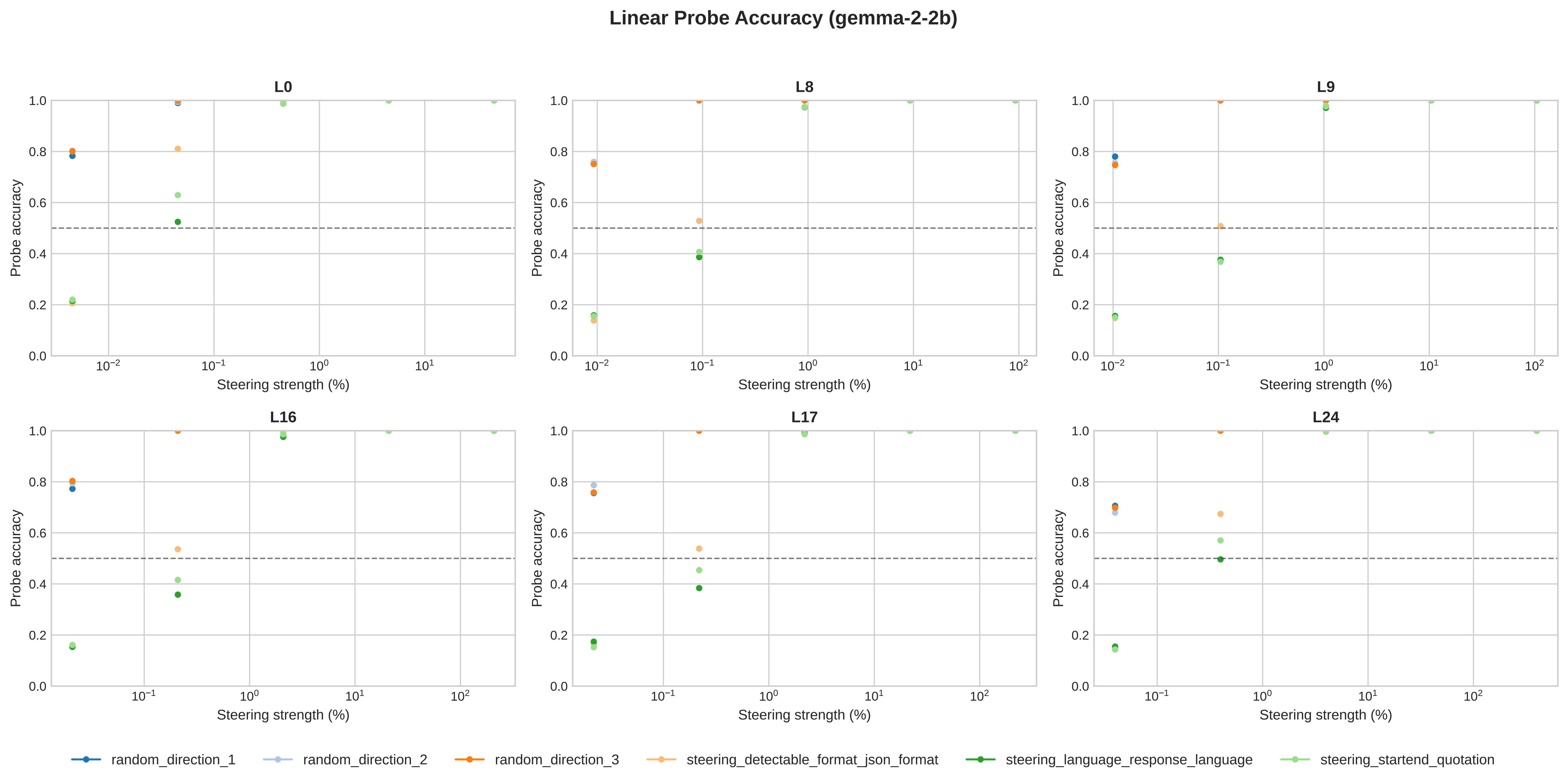
Linear Probe Accuracy (gemma-2-2b)
L0
L8
L9
L16
L17
L24
Probe accuracy
Steering strength (%)
0.0
0.2
0.4
0.6
0.8
1.0
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
random_direction_1
random_direction_2
random_direction_3
steering_detectable_format_json_format
steering_language_response_language
steering_startend_quotation