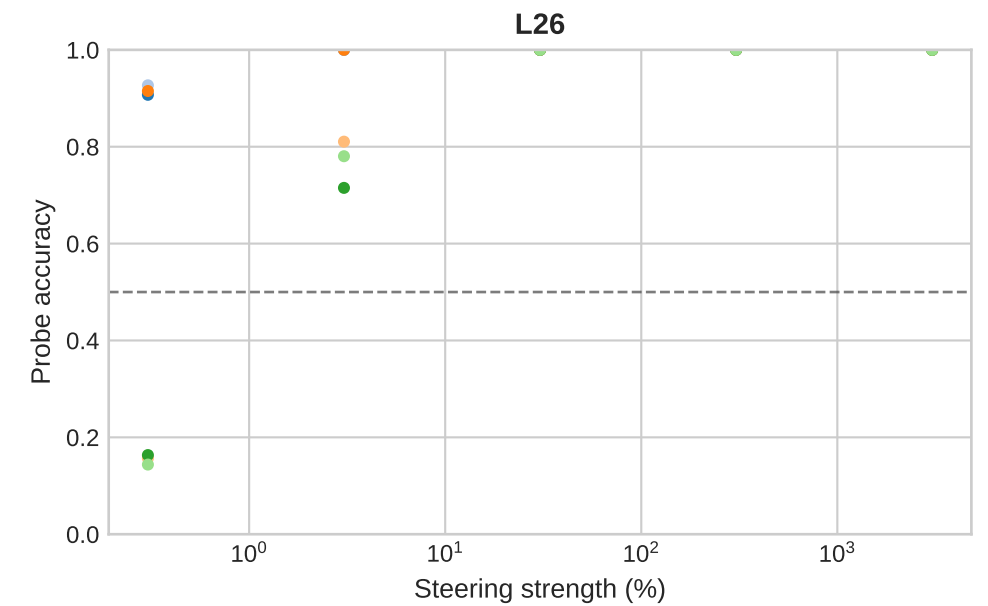
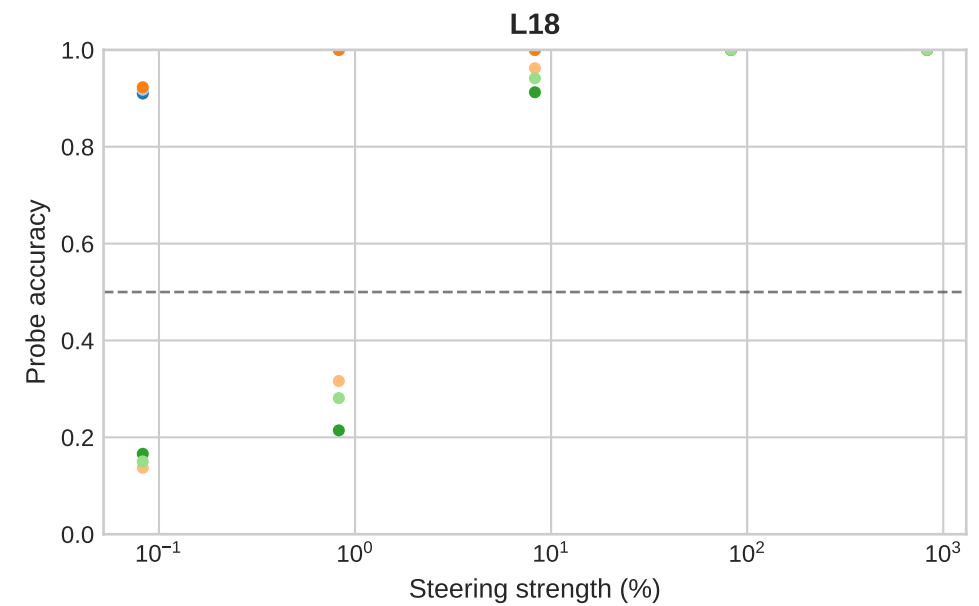
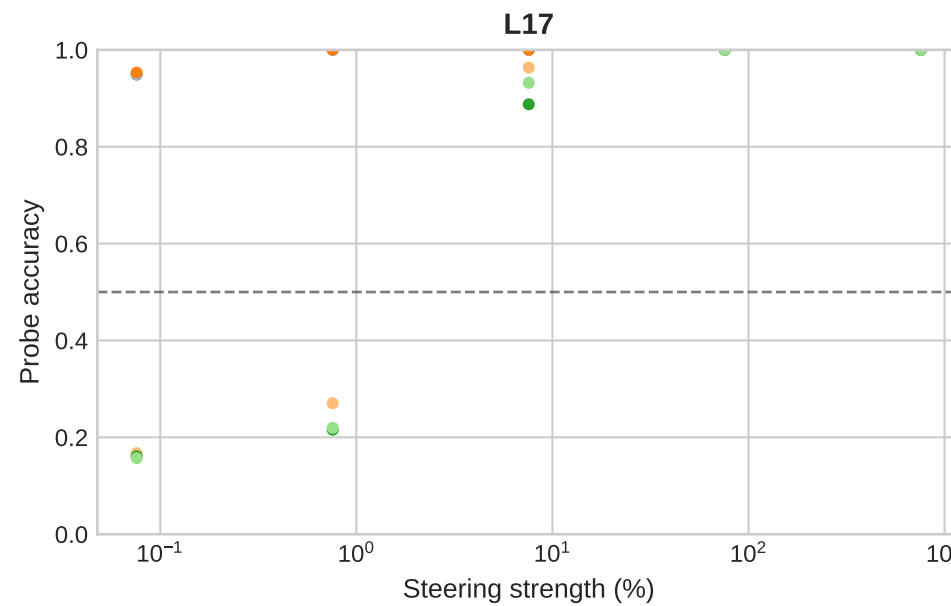
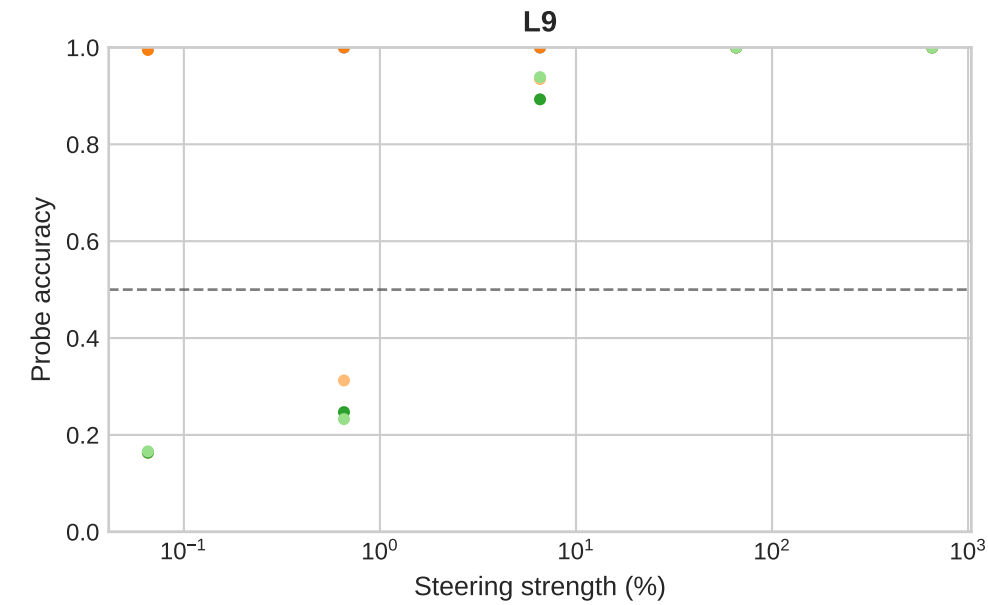
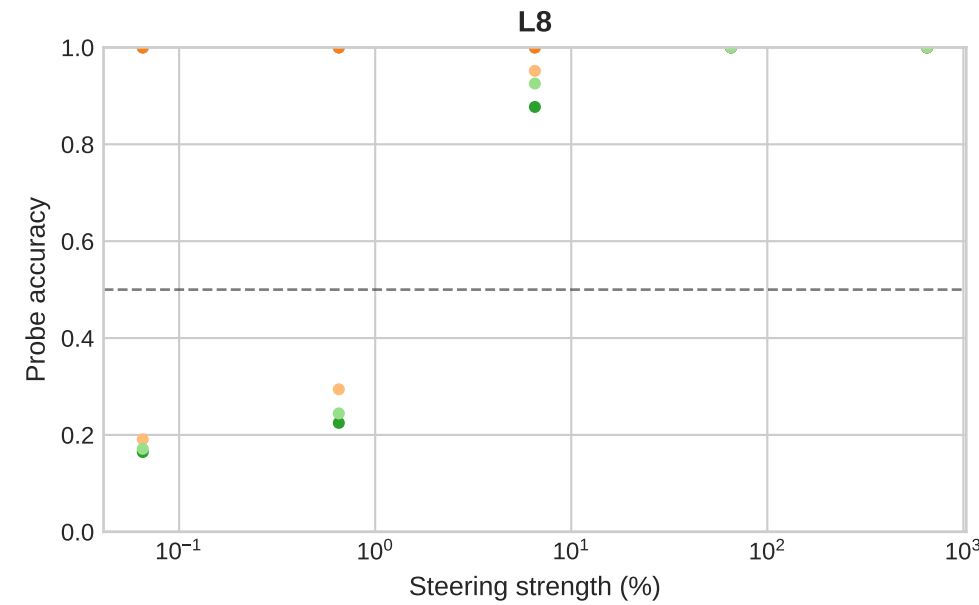
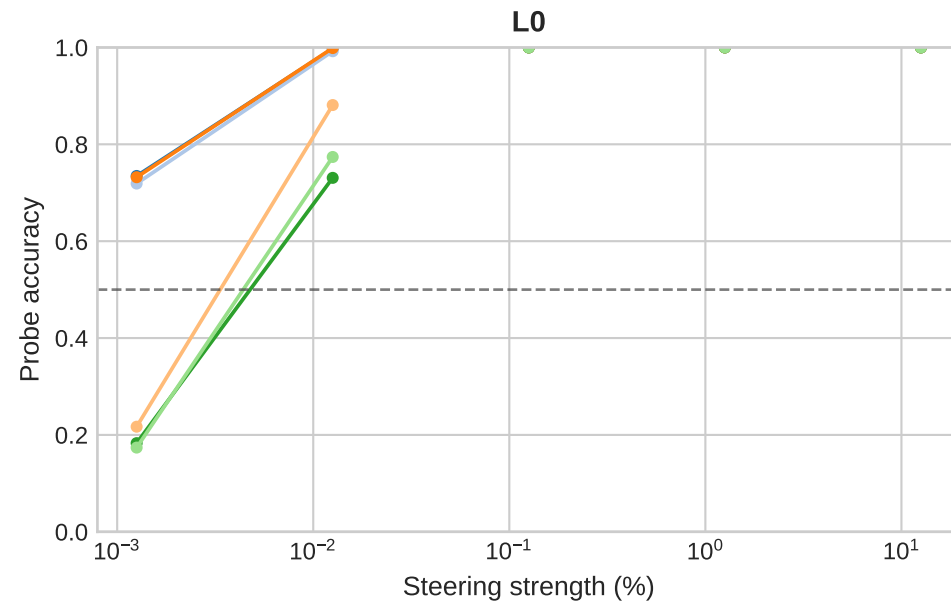


Linear Probe Accuracy (Qwen3-1.7B)



—●— random_direction_1
 —●— random_direction_2
 —●— random_direction_3
 —●— steering_detectable_format_json_format
 —●— steering_language_response_language
 —●— steering_startend_quotation