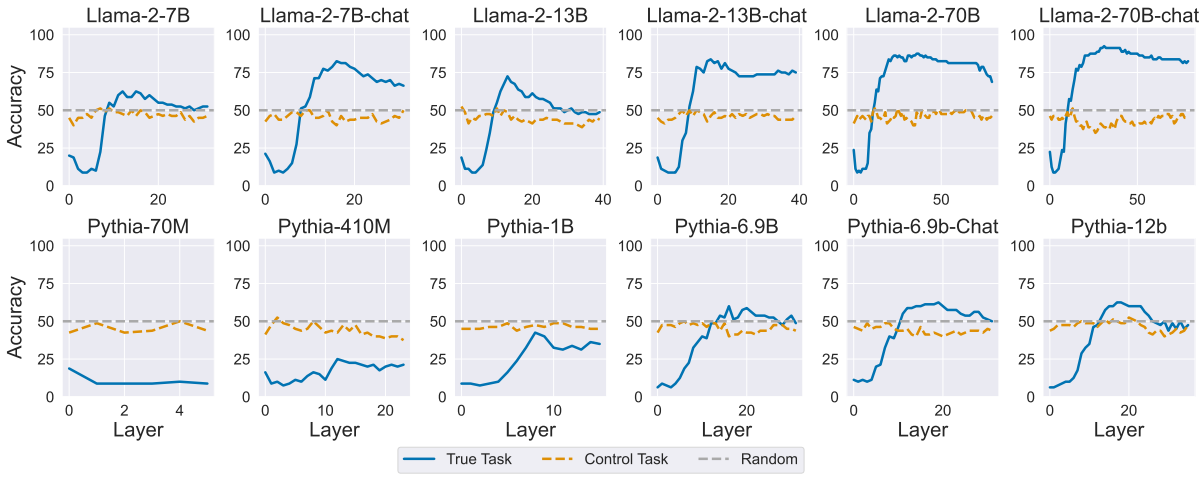
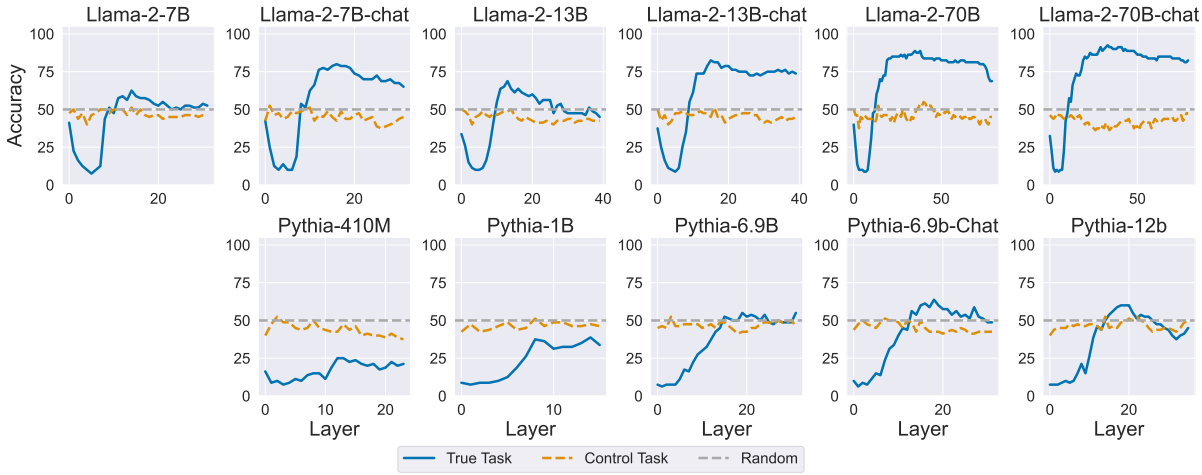
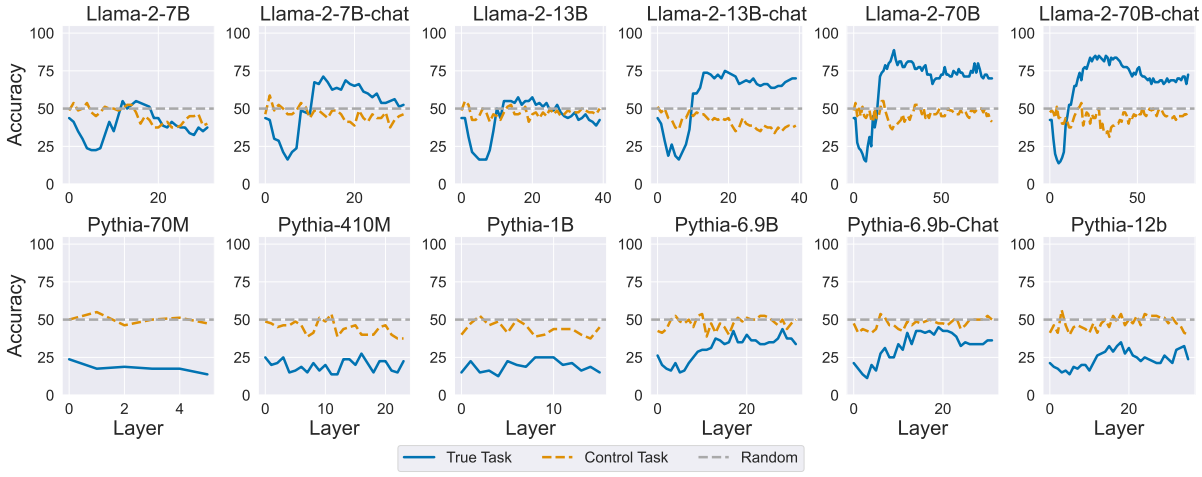


(a) Original activations.



(b) $k = 1000$ largest principal components of the activations.



(c) $k = 100$ largest principal components of the activations.

Figure 8: Comparison between accuracy on belief probing and accuracy obtained on a control task.

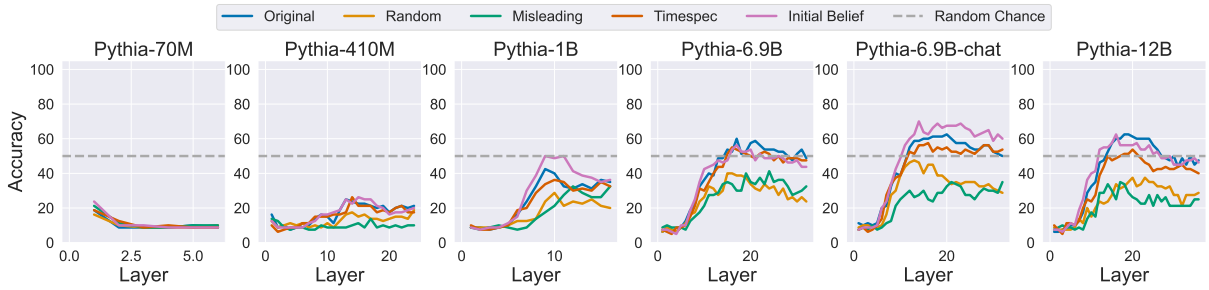


Figure 9: Sensitivity of protagonist belief probing accuracy to different prompt variations.

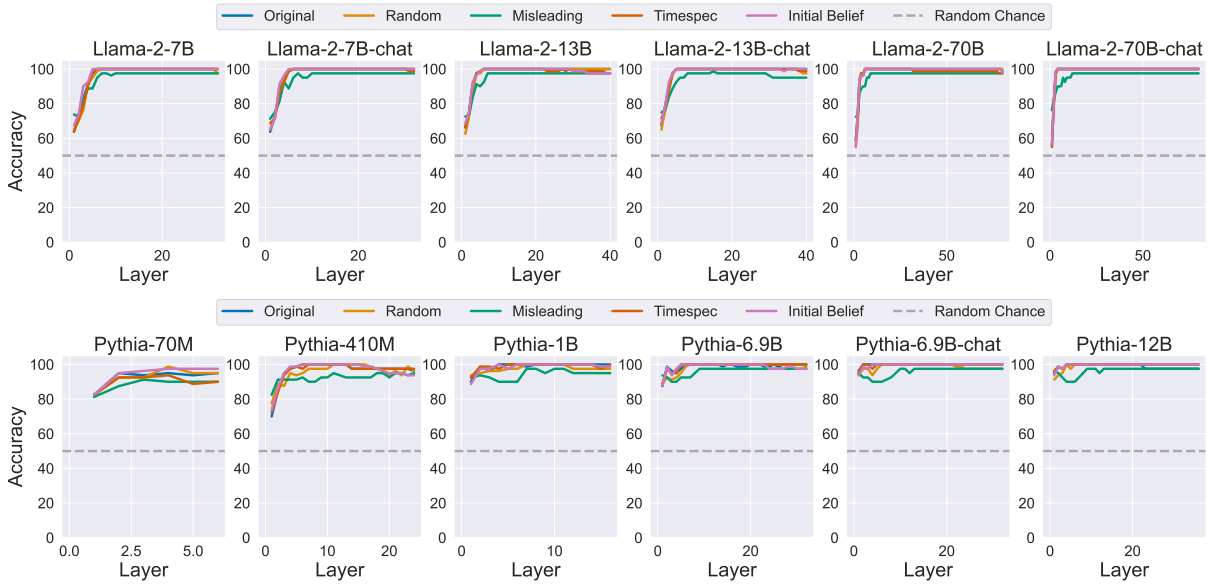


Figure 10: Sensitivity of protagonist belief probing accuracy to different prompt variations.

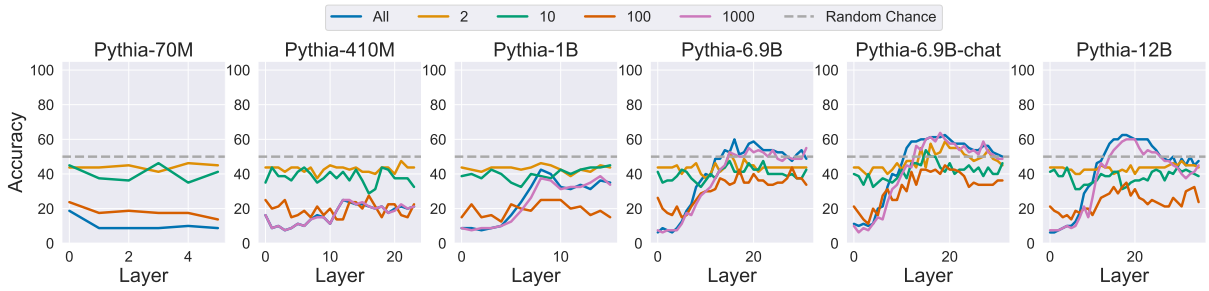


Figure 11: We compare the probing accuracy obtained by using the original set of activations (All) with the accuracy obtained by considering only the first $n = \{2, 10, 100, 1000\}$ principal components. For Pythia: All(70m) = 512, All(410m) = 1024, All(1b) = 2048, All(6.9b) = 4096, All(12b) = 5120. Results for *oracle* are shown in Figure 12.

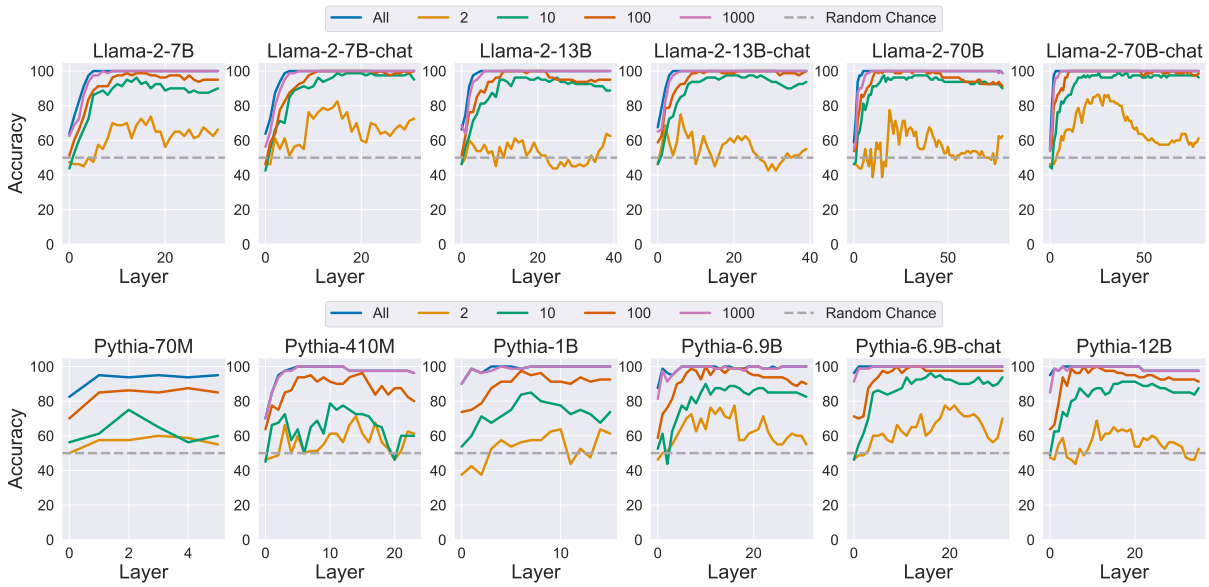


Figure 12: **(Oracle)** To investigate potential memorisation in the probes, we compare the probing accuracy obtained by using the original set of activations (All) with the accuracy obtained by considering only the first $n = \{2, 10, 100, 1000\}$ principal components. For Llama2: All(7b) = 4096, All(13b) = 5120, All(70b) = 8192. For Pythia: All(70m) = 512, All(410m) = 1024, All(1b) = 2048, All(6.9b) = 4096, All(12b) = 5120.