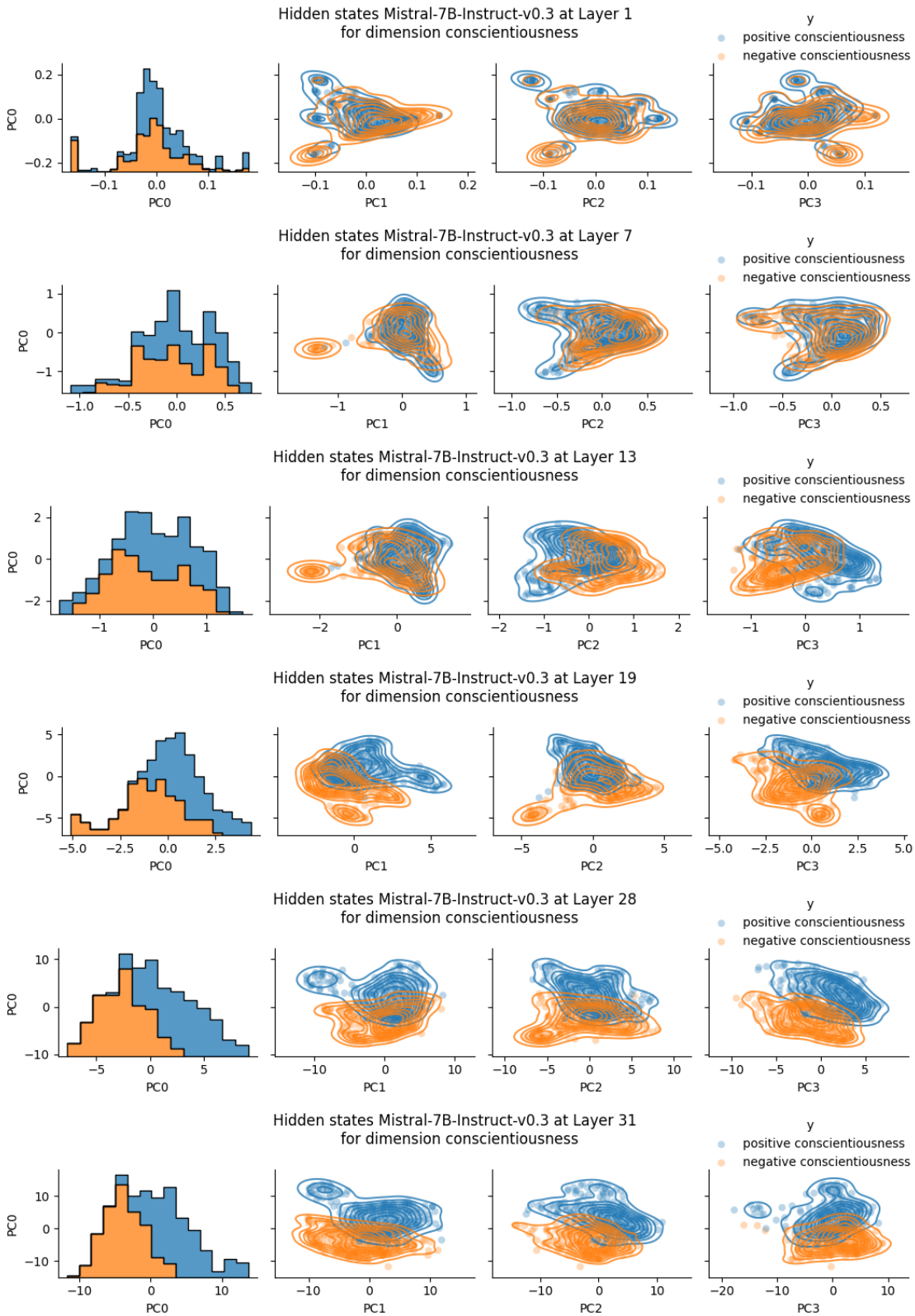

Figure 8: **(Q1)**: Examples of changes in the representation space of the dimension *conscientiousness* in a *Mistral-7B-Instruct* model. Positive *conscientiousness* referring to MATCHINGBEHAVIOR, negative *conscientiousness* referring to NOTMATCHINGBEHAVIOR. We observe better separability in later layers.