



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