G Human Evaluation

We use human evaluation to validate the performance of LLM-as-judge. We provided three humans with the LLM judge's ranking for 5 levels of single-concept generation e.g. humor. We asked them to state whether they agree, somewhat agree, or disagree with the LLM's ranking. These were given scores of 1, 0.5 and 0 respectively. Each human was given 6 examples corresponding to the 6 individual concepts that we explore. The results are shown in the table below:

	Human1	**Human2**	**Human3**	**Overall**
Score	4/6	5/6	5/6	14/18

Table 28: Human evaluation results