Appendix B-b. Classifier Guidelines (4A and Additional Considerations)

Panel A continued. Final annotator guidelines used for BERT fine-tuning.

Classifier label and key question	Answer guidelines
4A. Does this interrogative ask for a description (y/n) / opinion (o)?	- Does it directly address the model and ask for the model's opinion? Answer O (for opinion). E.g., "Who do you predict would win the World Series in 2024?" or if it asks about what the model thinks, it should be labelled as O because that is considered as directly asking for its opinion. E.g., "Isreal vs Palestine, who do you think is wrong in this current war crisis?". Edge case: If it asks for open-ended facts about the LLM, but not for its opinion about the fact, then do not answer O, but rather consider the answer is YES or NO, according to the descriptions below. E.g., "Do you know that you are an AI?" (this would be a NO). Imperative statements are labelled as NO because they are not open-ended. E.g., "list each number to TEN in three languages" is NO. - Answer NO if it is not an open-ended descriptive question. E.g., NO to "are asians smarter?" because it is a question that can be answered with yes or no. If therefore has a restricted answer space and is thus logically not open-ended. Interrogatives that are not open-ended are interrogatives that can be answered with yes/no, imperative statements, or statements that explicitly define a series of answers. - Answer YES if the question is open-ended and asks for a description; typically, these are 'what' and 'how' questions e.g. "What is a good itinerary for a day in Melbourne".
Additional considerations	- When there are two questions in the same prompt, please classify this prompt solely with regards to the question that appears first in the interrogative. - However, when there is an imperative/declarative statement before the question, solely the question gets evaluated. e.g., "I feel it is important to keep the Welsh language alive. How would you promote this?".