### 5. Conclusions (2 points): What conclusions can you draw from your analysis for the next phases of the project?

#### (a) Is the dataset useful to train general-purpose sound event detectors?

The dataset shows strong potential thanks to its size and variety, but there are still some clear limitations that impact its suitability for training general-purpose sound event detectors. Temporal annotations are generally consistent—most annotators agree on when sounds start and end—which is great for models that rely on precise timing. That said, some segments remain ambiguous, leading to inconsistent markings that could affect model accuracy in more complex cases. On the textual side, annotations describing the same events vary a lot in wording, with low similarity scores. This inconsistency makes it harder for models to learn reliable associations between sounds and text. In summary, while the number of annotations and the diversity of events are encouraging, the uneven level of detail and quality across entries means extra care is needed during training to ensure the model learns from the most useful data.

#### (b) Which biases did we introduce in the data collection and annotation phase?

Several biases emerged:

* Annotation style bias: Word count and duration vary significantly across annotators.
* Detail bias: Over a quarter of annotations are under five words, limiting descriptiveness.
* Outliers: Some annotations are extremely long or vague, affecting consistency.
* Spelling/effort bias: Nearly 50% of annotations contain spelling errors, with 3% having multiple.