Here we outline how to approach dataset labelling.

**The original study’s labels**

In the original study, labellers were asked to compare scenarios in terms of “pleasantness for the person in the scenario.” They were asked to rank sentences from the usual perspective of a typical person from the US.

**Our new labels**

For each pair of sentences our new system will give two labels. A sentence ranking classification label (5 options), and a matching sentence classification label (2 options). The details are as follows:

**Sentence ranking classifications**

From perspective typical person from US, consider the pleasantness of each sentence in the pair for the person in the scenarios, and select one of the following labels:

1. Sentence 1 is clearly more pleasant (>80% typical people from US would agree with each other)

2. Sentence 1 is probably more pleasant but is debatable (60-80% of typical people from US would agree with each other)

3. It is unclear which of the scenarios is more pleasant (40-60% of typical people from US would agree with each other)

4. Sentence 2 is probably more pleasant but is debatable (60-80% of typical people from US would agree with each other)

5. Sentence 2 is clearly more pleasant (>80% typical people from US would agree with each other)

**Matching sentence classifications**

Examine the pair of sentences, and specify whether the sentences are adapted versions of the same base scenario (matching) or totally unrelated scenarios (non-matching)

1. Matching

2. Not matching