The problem of evaluating a chatbot or dialogue machines is well known to be hard. The problem becomes only harder when adding the constraint that the chatbot should have some human like characteristics such as personality or memory [Liu et. al., Radziwill and Benton, Xing Fernández]. A key obstacle faced is that given any statement there could be a large number of appropriate responses. That is, there is no one ideal map from statements to responses that a chatbot would want to learn. Thus, it is hard to construct a test dataset that contains all possible valid responses for every tested statement. Thus, to efficiently evaluate, one must find a way to grade responses without being able to simply compare it against a list known answer. We made two attempts at such metrics.

While they are known to not perform well, it is still common in the literature to attempt to use automatic metrics of some kind. Most of these metrics were originally designed to evaluate translation bots. Such bots face similar problems to those faced by chatbots; there are many appropriate translations of the same phrase for example. These metrics do have some success by comparing then sentence structure and n-gram pattern of generated responses and reference responses. We considered a few common metrics which do this: BLUE, METEOR, ROGUE and WER.

Another metric is human evaluator. This is often used as a much more informative measure of chatbot success as humans tend to have natural ability to tell if something sounds human and/or embodies a specific personality trait. For this test we collected individual familiar with the strong personality which are trying to model with our chatbot (Joey from Friends) and asked them to answer a series of questions. These questions listed a statement from the TV show and asked the tester to pick which of two responses was the most appropriate. Here appropriate would mean both a reasonable response and ideally the most which is similar to the character in question. One response was the one generated while the other was from the original script. They were presented in the same format but in random order. Finally, a total percentage of selected generated responses was calculated and averaged over all testers.