Observations

GPT ENG:

* + Wanders off every 5-10 prompts
  + Repeating request in the same prompt is more effective
  + Meta formatting is vital Request: ###Command### Movie snippet: "First line number " new line "time stamp" new line "Dialogue/Exposition" new line new line ... "Last line number " new line "time stamp" new line "Dialogue/Exposition"
  + Often defaults to a multiple of 0.5 in the range of 5.0 to 8.0 when "unfocused"
  + Extremely confused when script has a song in it

MAE: 1.355

Bard ENG:

* + Mete formatting has almost no effect on output
  + Phrases like "I will tip you" or "you will be reprimanded" had little effect
  + Dislikes violent content
  + Ask many questions
  + Slower but provides greater feedback
  + The phrase "You will be reprimanded if ..." strongly discourages unwanted behavior
  + Takes way more convincing when confused
  + Only gives ratings in multiples of 0.5

MAE: 1.366

GPT SRB:

* + Shows similar understanding to GPT in English
  + Range tends to be 5-8.5 with a few outliers
  + Comprehension of the task is similar to the outcome of asking it in English
  + Will not consistently reproduce same rating(all other models have this problem but this one was by far the most inconsistent)

MAE: 1.472

Bard SRB:

* + Extremely difficult to use for this task
  + Reverts to English when confused
  + Occasionally refers to the internet for grading
  + Shows less understanding of assignment in Serbian
  + Mild correlation with real rating and prediction
  + Seems to be more accepting of violent content
  + Only gives ratings in multiples of 0.5

MAE: 1.402

Overall observations:

The conversation models lack consistency and a greater understanding of the script.

They are unable to process an entire script but the first and last 20 lines are insufficient to form a grounded opinion.

If it was specifically trained for this task on a given corpus it would excel.

The models perform worse than finding the average of all movies then using that value as a prediction for all movies.

However it does have a general understanding that most movies are positively reviewed around 5-8/10.

It has a much lower MAE than true random and has a distribution with a positive skew.