# Grading Scheme

### Grade:

* 0 - 5, with 0 being the most incorrect and 5 being the most correct
* Classify the reason for wrong answers (To check if we need to add context information to the model)
  + “Context error”: GPT answered the question wrong because it did not understand the context
  + “Error”: GPT answered the question wrong while it understood the context

Observation

* *Context is very important and helps the LLM to avoid misunderstanding*
* **The LLM has some knowledge about Goal-oriented requirement engineering but it is not perfect.**
* The questions that require reasoning/application are typically problematic, e.g. The numbers in E.12
* When the context is clear from the question, extra context information does not help too much
* The LLM can give reasonably good examples of goal models in a different language
* **Our grading is consistent on the correct/wrong answer but not very consistent on the deduction of parietal marks**

Plots/Table

* Average score of each question for different generation/grader
  + Average of the entire test suite (Grade for LLM).
* Variation of the answers for each generation