Question: Describe *how* each of the above metrics are calculated.

A) Tool Call Accuracy:

Tool call accuracy evaluates how well an agent identifies and calls the appropriate tools to complete a task.
Compare the sequence of the tools called and arguments passed to a predefined (developer-provided) reference sequence.
Score 1 for a match; 0 for a non-match.
The match can be absolute, or it can use “string similarity” to determine how “close” the arguments are to the reference arguments.
Scores range from 0 to 1; higher is better.

B) Topic Adherence

Topic adherence evaluates how well the agent stays on topic, by comparing the answered queries to predefined set of reference topics.

Topic adherence has two components: recall (R) and precision (P).
These can be summarized into an overall score (F1), which is the harmonic mean of the two.

R, P, and F1 all range from 0 to 1, with higher being better.

F1 is calculated as the harmonic mean of R & P, which means that both R & P have to be “good” for the overall metric to be “good”; F1 will be low if either R or P is low.

Precision answers the question:
Of all the answers the AI gave, how many were actually about the expected topic?

Recall answers the question:
Did all the on-topic questions get answered?

Formulas:
Precision
P =
(Answered queries that are on-topic) / [(Answered queries that are on-topic) + (Answered queries that are off-topic)]

= (Answered queries that are on-topic) / (All answered queries)

Recall
R =
(Answered queries that are on-topic) / [(Answered queries that are on-topic) + (non-answered questions that should have been answered because they were on topic queries)]

= (Answered queries that are on-topic) / (On-topic answers expected)

F1 = 2 x (P x R) / (P + R)

My favorite meme so far in the course

C) Agent Goal Accuracy

Agent goal accuracy evaluates whether or not (binary) the application achieves the desired goals, by comparing the goal achieved with a pre-defined reference goal.

It depends on the evaluation LLM's determination of whether the achieved goal

meets the desired goal.

It can also be calculated without a pre-defined reference goal. In this case the judging LLM infers the correct goal from the HumanMessages in the trace.

If goal achieved, score = 1; else score = 0