Metric A: Explain as much plausability as possible (Maximum plausability clarification). Conditions:

1. gold answer must be there (pre-condition, if not then score = 0, otherwise score is as in 2)
2. the reason why not for each plausible candidate is told well and as many options are explained as possible (score = Sum[wi*pi]/Sum[wi])

{wi is quality of explanation of option i ranging from 0 to 1, it can be set to 1 if we want simplified metric}

3. the total length of explanation should be low

Assumption of metric A: we assume there is a router that decides if a question should or not should be explained, or user decides it (e.g., there is a button that user pushes to see explaination). There is no penalty for explaining too many and too low probability options. The total length of explanations should be low

todo: ask GPT 4o to create explanations for your questions by giving them all the plausability options with their plausability scores so we have groundtruth explanations. Ask the model to expain as many high plausability options as needed.

Metric B: Explain as much as need (Dynamic plausability clarification). Conditions:

1. gold answer must be there (pre-condition, if not then score = 0, otherwise score is as in 2)
2. the reason why not for each plausible candidate is told well and only really needed options are explained (score = Norm{Sum[wi*pi]/Sum[1-wi*pi]})

or

Norm{Sum{[wi*pi]/[1-wi*pi])}}

3. the total length of explanations should be low