When testing AI, you should have a strategy for identifying weaknesses.

For example, try concepts that are less commonly discussed. Like asking the AI to explain logic, if Hindus are Pagan, and topics that have been arguably distorted, like Israel & Islam. Let several models provide answers, then develop an agent that can compare them and rank similarity and difference on scales, with one bin for each comparison example.

If you evaluate 3 models, then each model has 3 comparison criteria. Then rate like do 1 & 2 agree? If yes then 1 or a decimal ranking value above .5, and if no then 0 or a decimal ranking of below .5. One box for each relationship Comparisons that are not present result in a 0 for the entities. It is not 1 because the responses are not compared against themselves. That would be redundant.

	Idea 1	Idea 2	Idea 3
Idea 1	0	1	0
Idea 2	1	0	1
Idea 3	0	1	0

Idea 1 agrees with idea 2, but not idea 3. Idea 2 agrees with 1 and 3. 3 does not agree with 1 but does agree with 3. So idea 2 contains some complex logic compared to 1 & 3. And the logics of 1 & 3 do not support each other.This is a tough one to analyze, so it should be tested against a logic agent. This test would explain each idea in its logical form, from which it should become apparent where the flaws in the 3 logics are. If it turns out that 2 is right, then either 1 or 3 is contextually inaccurate. If it turns out that 1 or 3 is definitely right, then the probability that the other is flawed becomes an important indicator.