For dataset:add noise to questions ie out of curriculum questions like 5 more - 2 hard,2 medium,1 easy to identify
You might prefer to test 1 by one
Add columns one by one
An evaluation of the Math.ai agent on various metrics considered important for its real world deployment:

| Test topic | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Groundedness in curriculum | | | | | | | | | | | | | | | | | | | |
| Language Accessibility | | | | | | | | | | | | | | | | | | | |
| Accuracy | | | | | | | | | | | | | | | | | | | |
| Helpfulness | | | | | | | | | | | | | | | | | | | |
| Soundness of steps(no of sound steps up to where it vagued/totalexpected steps) | | | | | | | | | | | | | | | | | | | |
| adaptability | | | | | | | | | | | | | | | | | | | |
| Overall system performance(aggregate) | | | | | | | | | | | | | | | | | | | |

Meanings of test topics:

- **Groundedness in curriculum:** How well the agent's answers and explanations align with the specific educational curriculum it is designed to use.
- **Language Accessibility:** How clear and easy-to-understand the agent's language and explanations are for the user.
- **Accuracy:** The correctness of the agent's final answer and calculations.
- **Helpfulness:** The quality and utility of the agent's guidance and support provided to the user.
- **Soundness of steps:** Measures whether the intermediate steps in the solution are logically correct (calculated as the number of sound steps before an error divided by the total expected steps).

- **Adaptability:** The agent's ability to handle different problem types, varied user inputs, or unexpected situations.
- **Overall system performance (aggregate):** A combined score representing the agent's total performance across all the other key metrics.