Rethinking Math Benchmarks for LLMs using IRT

Jane Castleman*, Nimra Nadeem*, Tanvi Namjoshi*, Lydia T. Liu

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Key Research Questions:

(1) How robustly do current benchmarks estimate and rank LLM abilities for AIED use?

(2) How can we design benchmarks that remain effective as model abilities increase?

Background (IRT):

Item Response Theory: measures the latent abilities of test-takers and the difficulty and discriminability of questions

$$P(x_{i,j} = 1 \,|\, \theta_i, b_j, a_j) = \frac{1}{1 + exp[-a_j(\theta_i - b_j)]}$$

$x_{i,j} \rightarrow$ model i's response to item j

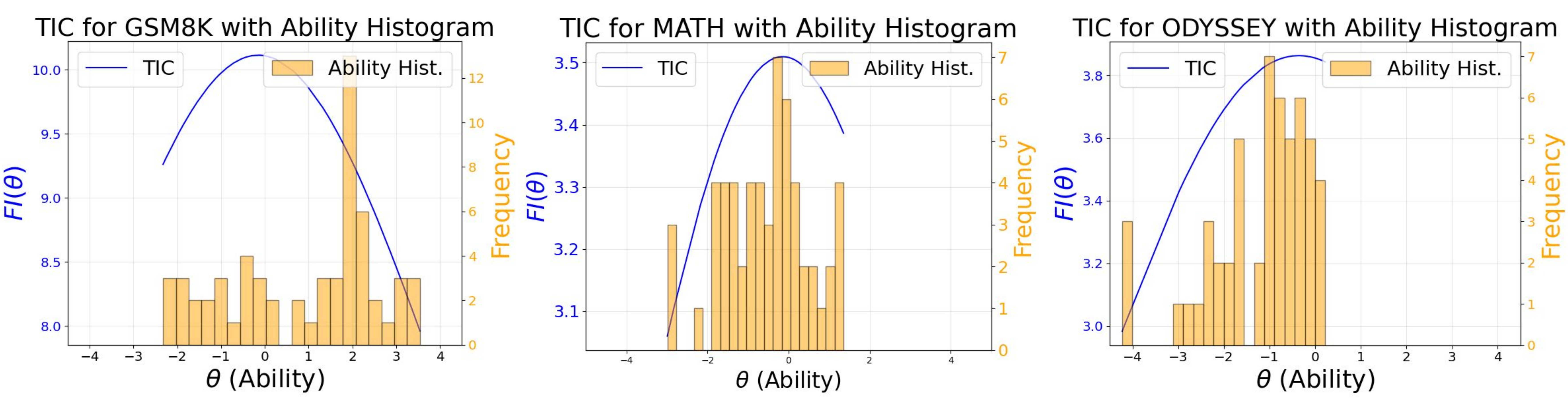
$\theta_i \rightarrow$ latent ability of model

$a_j \rightarrow$ item discrimination $b_j \rightarrow$ item difficulty

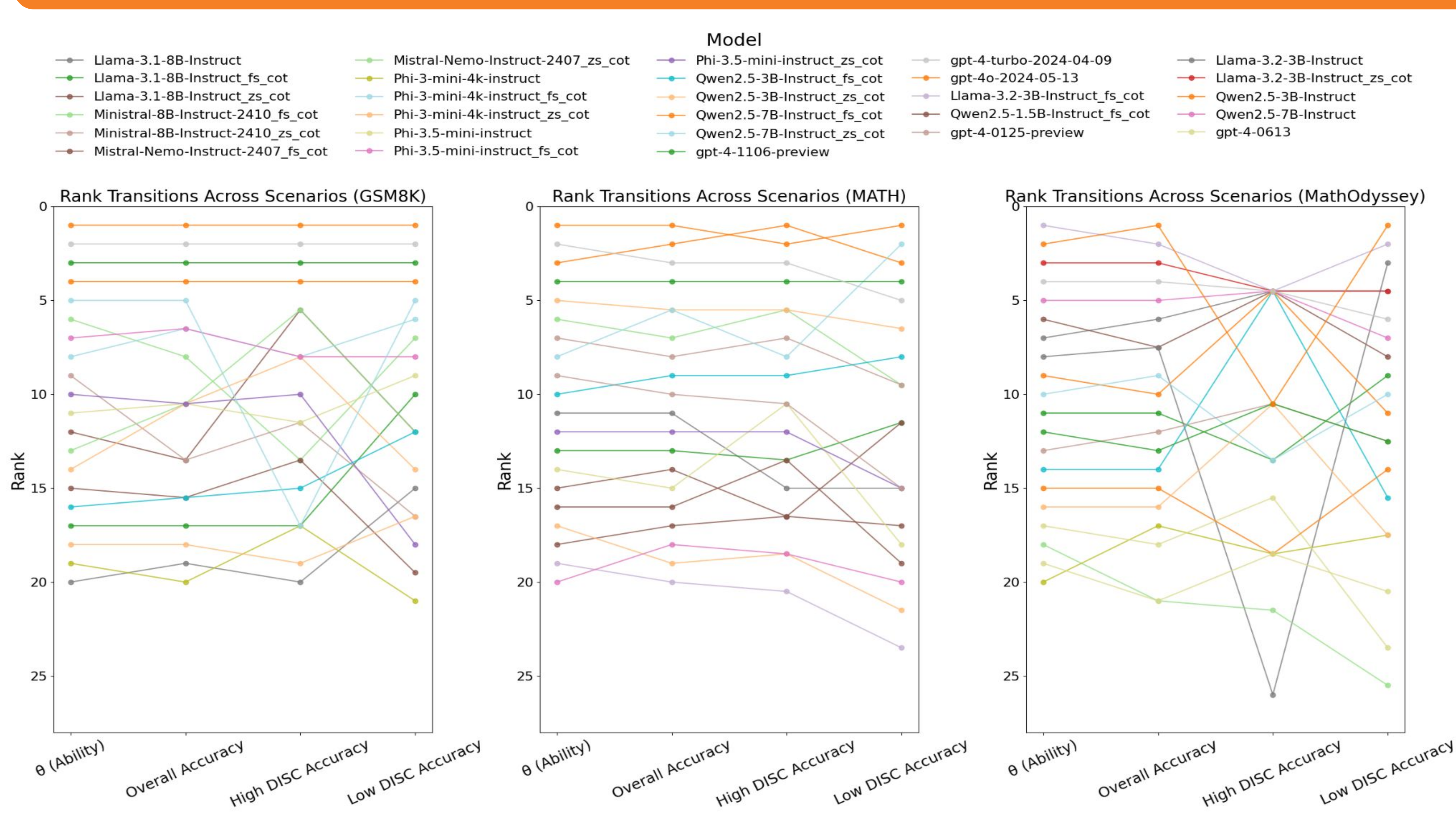
Methods:

- **Benchmark Datasets:** GSM8K, MATH, MathOdyssey
- **Test-taking population of LLMs:** Frontier models ranging from 0.5B to 1T parameters with different prompting strategies (CoT)
- **Model:** 2-parameter IRT using the `py-irt` python package

We find GSM8K and MathOdyssey provide limited information for the current range of SOTA models; MATH is the best-suited benchmark for today's abilities



Model rankings based on overall accuracy are unstable across subsets



- Model rankings change when using high discrimination questions
- Challenging for AIED practitioners to identify most capable models

Current benchmarks may be unreliable for assessing model abilities; IRT is a promising approach for finding highly discriminative questions