The Correct Answer Trap:
Pedagogically-grounded detection and feedback generation for hidden misconceptions

Moiz Imran
Department of Computer Science
University College London

Sahan Bulathwela
Centre for Artificial Intelligence
University College London

International Conference on Artificial Intelligence in Education 2026

The Bigger Picture

Where do the pieces fit in?

Intelligent Tutoring as a Society of Minds: An Agentic Approach

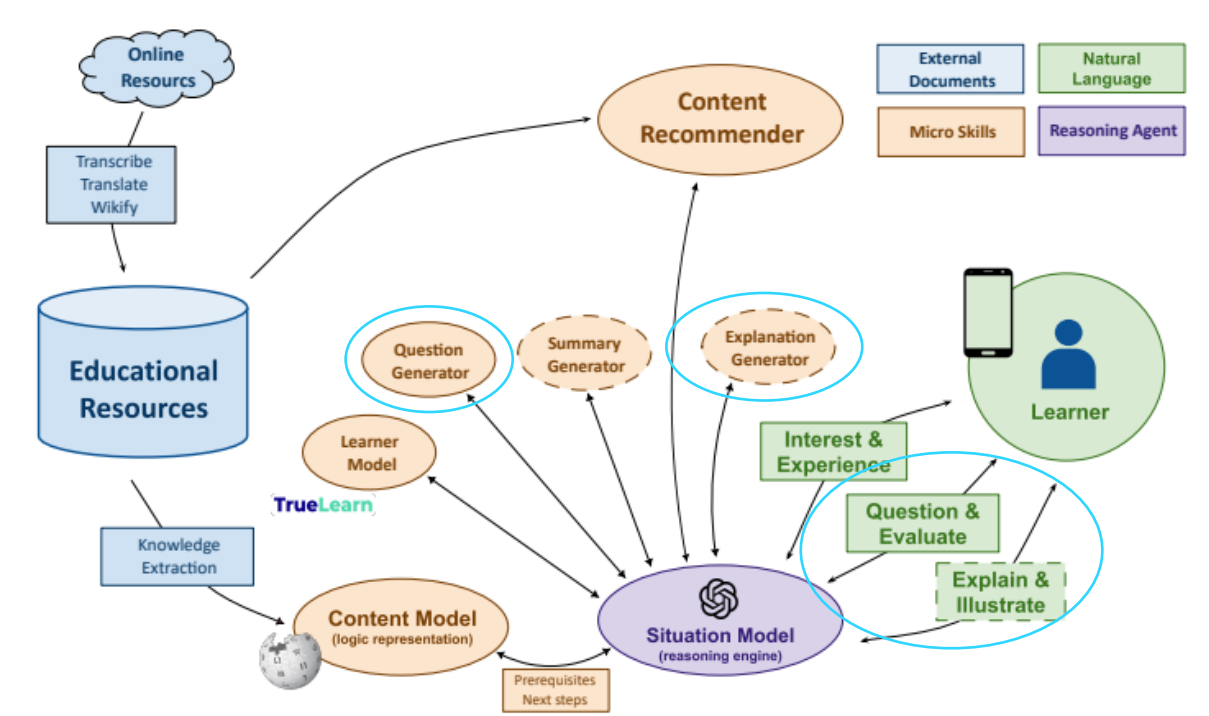


Fig. 3 The logical components of the TrueReason learning assistant consist of the blue parts that are learning resources, the orange components belong to the core AI assistant that entails 1) the domain model (that is mapped to Wikipedia topics), 2) the learner model, and 3) the pedagogical model, which includes the engagement model and the situation model that interacts with the learner and the green part represents the user interface. The components (e.g. micro skills) that are not implemented at present use dashed lines.

The Task

Misconception detection from short student explanations

- Precursor to automated feedback in AI tutoring systems

Given a student's answer and explanation, classify reasoning as:

- Correct answer + accurate reasoning
- Incorrect answer + flawed reasoning
- Correct answer, flawed reasoning

20,964 real responses from the Eedi diagnostic maths platform

5 models evaluated

- Fine-tuned classifiers
- Frontier LLMs

The Correct Answer Trap

2% of correct answers hide flawed reasoning.

Question: (−8) − (−5)

Student: "8 minus 5 is 3, since there are negatives, −3."

Correct answer, flawed reasoning.

- Rule works here but fails on (−5) − (−8)

Dataset and Evaluation

Labels (human-annotated):

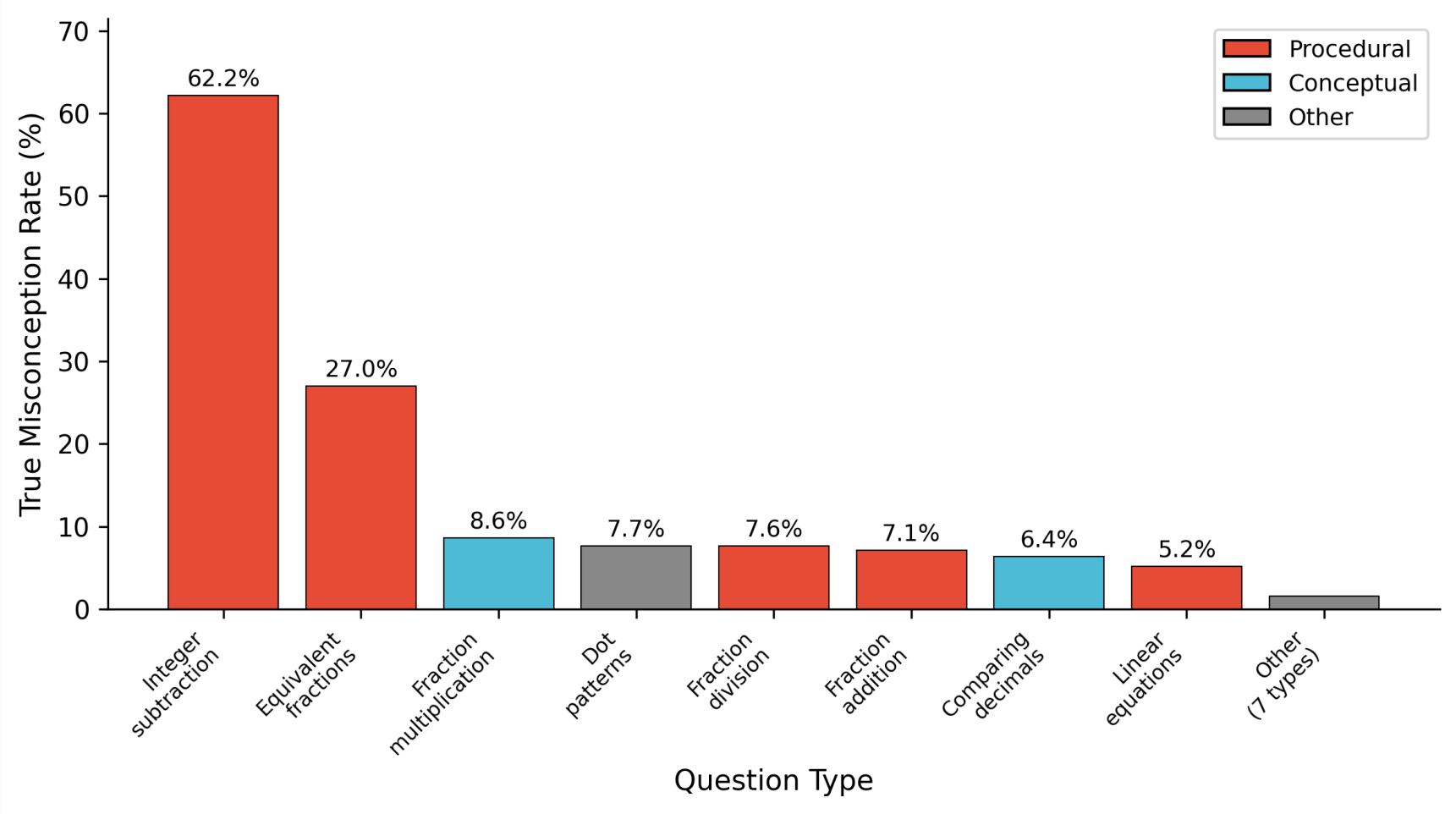
- True Correct (TC): correct answer, sound reasoning – 60%
- False Misconception (FM): wrong answer, flawed reasoning – 38%
- True Misconception (TM): correct answer, flawed reasoning – 2%

Scale: 343 TM cases total; 61 in held-out test set; 15 unique questions

Signal: average explanation ~16 words (short, noisy student text)

Finding 1: Failures Concentrate

71% of all TM cases occur in just 2 questions (out of 15)



Finding 2: Why These Questions?

Coincidental correctness paths

- A common flawed procedure happens to produce the correct answer for the specific values chosen

(−8) − (−5): “subtract then make negative” → gives −3 (correct)

- Same rule fails on (−5) − (−8)

A/10 = 9/15: mechanical cross-multiplication → gives A=6 (correct)

- Without understanding proportional reasoning

Barton (2018): “It should not be possible to answer correctly whilst still holding a misconception”

Finding 3: The Deployment Gap

Fine-tuned models (BERT, T5):

- Near-perfect at answer checking (FM recall ≥98%)
- Fail on reasoning assessment: TM recall 54–57%

Frontier LLM (Gemini 3 Flash):

- Best balance: 84% TM recall, 94% TC recall
- Improvement over fine-tuned significant ($p = 0.005$)

Open-source LLMs (Llama 70B, 8B):

- Similar TM recall (82–84%) but TC recall 55–68%
- High recall is cheap if you flag everyone

The Deployment Reality

At natural TM prevalence of 1.6%:

Even the best model (84% TM recall, 94% TC recall) produces:

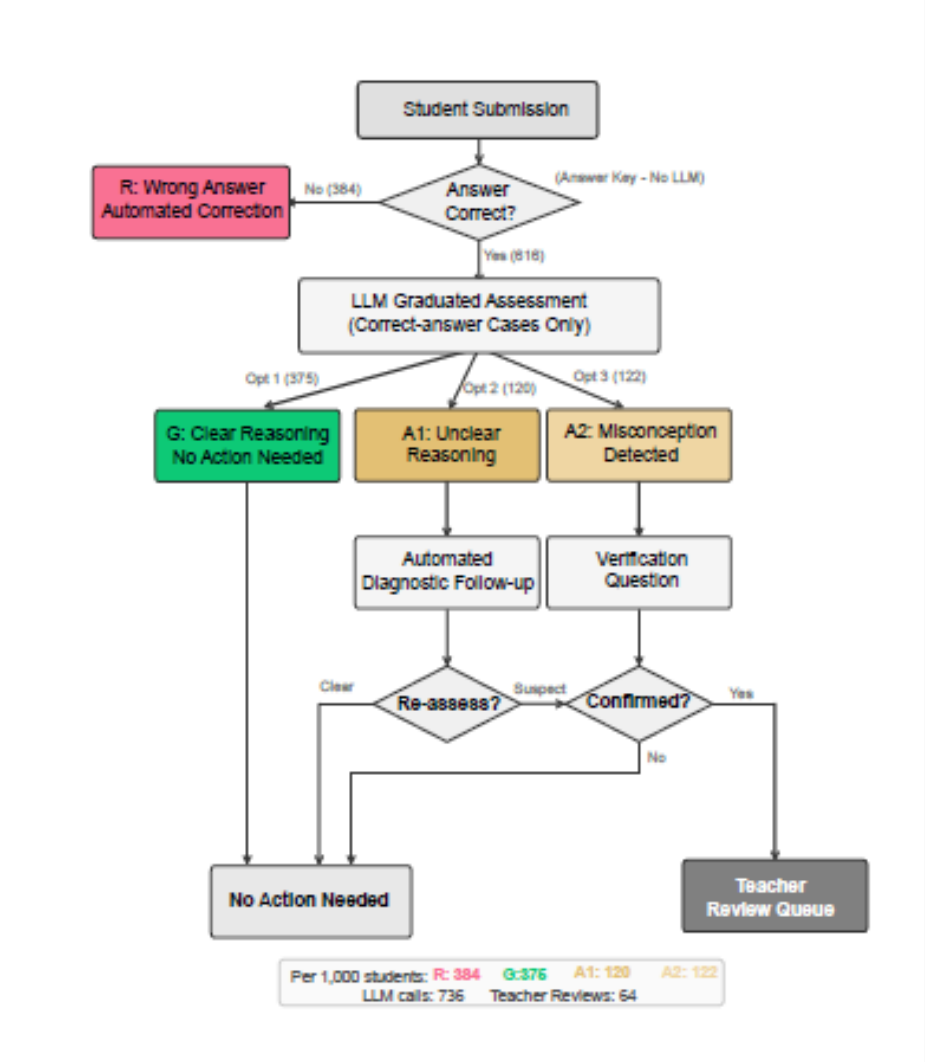
- 4.3 false alarms for every genuine detection
- Positive Predictive Value = 19%

For every 5 students flagged, only 1 has a misconception

- The other 4 were reasoning correctly

Accuracy is the wrong metric at this prevalence

The Deployment Reality



Implications

Question Design

- Check at design time: can a known misconception produce the correct answer for your chosen values?
- If yes — change the values
- Testable before deployment (Barton's diagnostic principle)

AI Deployment

- Stand-alone screening not viable at natural prevalence
- Risk-stratified: flag vulnerable questions, not all questions
- Follow-up probes more practical than binary classification

Conclusion

High accuracy masks failures precisely where intervention matters most.

The first fix is better question design, not bigger models.

Conference Paper
Monday June 29, 9am–10:15am, North 206
Code: github.com/Moiz-I/correct-answer-trap

Contact: m.bulathwela@ucl.ac.uk

UCL