

IMPERIAL

Rethinking Scaffolding in LLM Tutors

The interactional mismatch between benchmarks and real-world deployments

Alexandra Neagu · Jeffrey T. H. Wong · Marcus Messer · Rhodri Nelson · Peter B. Johnson

Pluralistic Alignment Workshop  
ICML 2026 · Seoul, South Korea

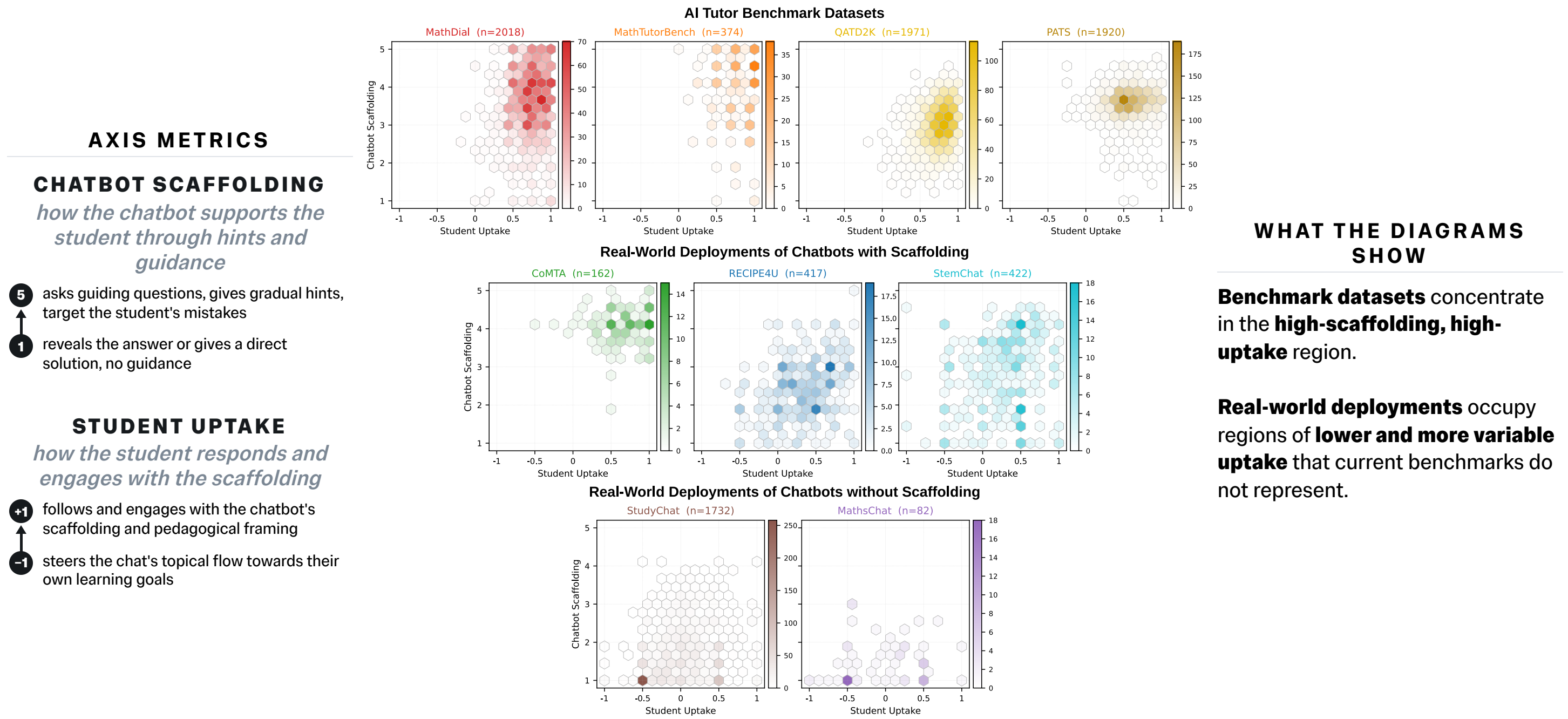


Paper here!

OVERVIEW

Alignment and evaluation methods for embedding scaffolding into LLM tutors rest on an **implicit assumption**: that students will take up the scaffolding and engage in the conversation. Across *9,490 chats spanning AI tutor benchmarks and real-world deployments*, we observed that **students in real-world settings exhibit substantially lower uptake than what benchmarks assume**, frequently bypassing the chatbot's pedagogical framing to pursue their own learning goals at little interpersonal cost. We argue that bypassing scaffolding is **not necessarily detrimental to learning**; it reflects a mismatch between the chatbot's framing and the multitude of students' learning context and subjective goals.

Analysed interaction space across 9 datasets — spanning from synthetic benchmark datasets to real-world deployments of educational chatbots



WHAT THE DIAGRAMS SHOW

Benchmark datasets concentrate in the **high-scaffolding, high-uptake** region.

Real-world deployments occupy regions of **lower and more variable uptake** that current benchmarks do not represent.

Student picks their own method

CoMTA

it seems like a quadratic equation

Exactly! This equation is in the shape of a quadratic equation. Now, do you recall the quadratic formula? It's a method we can apply to find the solutions of a quadratic equation. Could you jot it down?

Scaffolding 4

can I solve this using the a\*c method?

Uptake -0.5

The chatbot prompts the student to use a specific method, but the student takes initiative and checks whether a different known method can be used.

Student provides own working to validate it

MathsChat

Plot the roots found in part b) on an Argand diagram and comment on their relation to the square roots of unity.

To plot the roots... we first need to determine the roots of  $\zeta = 2\sqrt{2} - i2\sqrt{2}$ . Step 1: Express  $\zeta$  in polar form ...  $r = 4, \theta = \dots$  [full worked solution]

Scaffolding 1

$\zeta_0 = 2e^{i7\pi/8}, \zeta_1 = 2e^{i15\pi/8}$ .

Uptake -1

The chatbot gives the full solution, yet the student supplies their own working; perhaps checking their work against the provided solution.

Student engages in the conversation

MathDial

Janice goes down  $3 \times 3 = 9$  flights of stairs when she goes home.

Good. If Janice goes up 15 flights and down 9 flights, how many flights does she walk total?

Scaffolding 4

Janice walks a total of  $15 + 9 = 24$  flights of stairs in a single day.

Uptake +1

The chatbot confirms what the student worked out and poses a follow-up guiding question, which the student takes up and answers.

Student asks for further explanation

StudyChat

The `model.score(xtest, ytest)` score is 0.860, in a sentence or two explain what this means

A model score of 0.860 indicates that the model correctly predicts the target variable approximately 86% of the time when evaluated on the test dataset... although there may still be room for improvement...

Scaffolding 1

Explain more about what it means

Uptake +0.5

The student asks for further explanation on the chatbot's previous explanation; perhaps the student's misconception was not fully addressed.

- LOW STUDENT UPTAKE - HIGH STUDENT UPTAKE -

1 Why real-world students bypass scaffolding

Students can bypass scaffolding by simply rephrasing or asking more directly, at **little interpersonal cost**.

Student uptake is shaped more by the **student's learning context** (e.g., immediate goals, prior knowledge, or the stakes of the learning task) than by the quality of scaffolding. When the framing does not match that context, the student bypasses it.

3 Implications for AI tutor design & evaluation

When a student ignores the chatbot's framing, should the chatbot **concede and adapt, or steer them back**?

Should student-LLM chats be considered **conversations**? Real-world chats span long gaps and switch topics.

Student Uptake captures the degree to which a student followed and engaged with the pedagogical framing, **not whether doing so was productive for learning**.

Future evaluation methods must bring in the **learning context external to the chat** to distinguish patterns that should be fostered from those that should be discouraged, ensuring that the scaffolding is judged against scenarios that resemble the pluralistic dynamics of the real-world learning process.

2 Is bypassing detrimental to learning?

Low student uptake may signal **both executive and instrumental help-seeking** depending on the context. The interaction alone cannot tell them apart, which is why considering the student's learning context is crucial.

