# What state-of-the-art AI learning software looks like in 2026

Peabody blog · Learning science · updated 2026-11-13

Not a smarter chatbot. What the 2025 and 2026 trials say the best adaptive learning software does, layer by layer, and which parts matter most for a child.

The best AI learning software in 2026 is not a smarter chatbot. The trials published this year say an AI tutor helps most at one moment, right after a mistake, and only inside a structured practice system that already picks the right difficulty and brings things back at the right time. Built well, the AI reads what the child did, steps in when they get something wrong, writes and checks the next question, and hands the hard cases to a grown-up. A statistical model, not a language model, keeps track of what the child knows. Peabody (peabody.kids), free through the end of 2026, is built along these lines, with no language model in its practice at all. Here is what the evidence says and what a system built on it would look like.

## What the newest trials found

Three studies from 2025 and 2026 changed the picture.

An AI tutor one click away goes mostly unused. The economist Philip Oreopoulos and a colleague ran a two-year randomized trial in 18 Tennessee middle schools with Khan Academy's Khanmigo tutor. Maths rose by about 0.06 to 0.08 standard deviations a year, about what Khan Academy practice gives with no AI. Nearly every student tried the tutor, but the median student used it in only 17% of the practice sessions where they made a mistake. The working paper calls engagement the key constraint.

The AI helps most right after a mistake, inside a structure. A second study by the same group, with more than 6,000 middle schoolers, crossed AI help with mastery practice. The AI's clearest benefit was on the next attempt after a wrong answer. Mastery on its own did not lift scores on a later test. AI and mastery together did best, by a small margin. The authors conclude that AI tutoring adds value only with structure that turns mistakes into learning moments.

An AI that hands over answers does harm. In a study of about 1,000 high-school students in Turkey, published in PNAS, plain GPT-4 raised scores on practice problems by 48%. When the AI was taken away for the exam, those students did up to 17% worse than students who never had it. A version built to give hints instead of answers avoided most of the harm. The OECD's Digital Education Outlook 2026 draws the same line: output with AI is not learning.

The positive results share a pattern: a bounded task, a solution the system already knows, scaffolding the software enforces, and often a person in the loop. Our roundup of the 2024 to

2026 trials has the earlier studies.

## The system, in nine layers

If you designed learning software from that evidence, it would have nine parts:

- Layer: 1. Evidence · What it does: Turns taps, speech, handwriting and timing into a record of what the child did, including which wrong idea a wrong answer shows · What it is not: A score
- Layer: 2. Learner model · What it does: Estimates what the child knows, with an error bar, and what they are likely to forget · What it is not: A language model guessing
- Layer: 3. Content · What it does: Writes the next question, then checks it by another route (a solver for maths, a word list for early reading) before a child sees it · What it is not: Unchecked generation
- Layer: 4. Choosing what's next · What it does: Aims for about 80 to 85% success on new material, brings old skills back on a schedule, and changes approach when a child is stuck · What it is not: One fixed path
- Layer: 5. Experience · What it does: Puts the learning inside something the child wants to do · What it is not: A quiz with a game painted on
- Layer: 6. Tutor moves · What it does: Steps in at a mistake, one step at a time, never giving the answer before a real try · What it is not: An open chat window
- Layer: 7. Grown-ups · What it does: Tells a parent or teacher one specific thing to try, and flags what the software can't fix · What it is not: A dashboard of charts
- Layer: 8. Measurement · What it does: Judges every change by a later question the child answers alone, never with help on screen · What it is not: Time on app
- Layer: 9. Safety · What it does: Asks for no personal details, keeps raw voice on the device, and uses no guilt to bring a child back · What it is not: An afterthought

Two of these deserve a closer look.

The learner model should be statistics, not a chatbot. A 2026 benchmark asked large open language models to predict a student's next answer from their history. On their own they barely beat "always guess right". The models that lead the field are ordinary statistical trackers, sometimes with a language model describing the questions. Our knowledge tracing explainer covers the options.

Measure what the child can do later, alone. A child who succeeds only while help is on screen hasn't learned the skill yet. The best systems check a skill again days or weeks later, with a new question and no help, and steer by that number.

## A twelve-minute session

Here is what that looks like for a seven-year-old. She opens the app and her game is already playing. A gate asks 13 take away 5 and she says 7, a slip she has made before. The game doesn't stop. Two screens later she counts eggs in a nest and gets the same kind of question right with no hint. She reads a page aloud in a story written at her level. She gets three questions on a new skill wrong in a row, so the software switches to blocks on screen and asks one question: "How many tens do you see?" Later it asks a question from Tuesday, with no help offered, and she gets it right. At bedtime her dad gets one line: what she is solid on, what is new, and a two-minute thing to try with spoons.

Nothing in that session needed a chat window.

## Where Peabody stands

We compared Peabody's practice code against this design. Most of it is already there, written as ordinary code rather than AI:

- Every wrong box in maths facts and phonics stands for a named mistake, and the line after a miss says which one ("5 take away 3 is 2, not 8: you added").
- A retired question's skill comes back with a fresh question about 3 days, 2 weeks and 2 months later.
- A child stuck after ten tries gets the question another way, then an easier rung, then a worked example. Our post on how adaptive learning knows where a child is stuck explains how.
- Questions are kept rare in the games, and the learning dial starts on Light.

What's left: number sense and sentence questions don't name their wrong answers yet; a skill a child once passed doesn't re-open on its own when they start missing it; and no estimate carries an error bar. Each is written up as a build task. The one thing we don't plan to add to practice is a language model. Nothing in the evidence says a child needs one there.

You can start free in any browser, or pick a story from the library and watch which questions come back when.

## Sources


- Oreopoulos and Low (2026), One Click Away: AI Tutoring with Khanmigo in a Two-Year School Experiment, NBER.
- Oreopoulos, Liut, Sungu and Low (2026), Making AI Tutoring Productive, NBER.
- Bastani et al. (2025), Generative AI without guardrails can harm learning, PNAS.
- OECD (2026), Digital Education Outlook 2026.
- FoundationalASSIST (2026), a knowledge tracing benchmark for language models, arXiv.

Free at peabody.kids/blog/what-state-of-the-art-ai-learning-software-looks-like