Judging AI learning software: a nine-layer checklist

Peabody guide · updated 2026-11-13

Nine questions that sort real adaptive learning from a chatbot with a lesson plan, with what a good answer sounds like and the 2025 and 2026 evidence behind each.

Good AI learning software is a structured practice system with AI doing a few narrow jobs inside it: reading what the child did, stepping in after a mistake, writing and checking the next question, and passing hard cases to a grown-up. This checklist turns that into nine questions you can ask of any app, one per layer, with what a good answer sounds like. It comes from the randomized trials of 2025 and 2026, which found that an AI tutor added little when it sat in a chat window, did harm when it handed over answers, and helped most right after a mistake inside a structure. Peabody (peabody.kids), free through the end of 2026, answers each question at the end of its row.

How to use it: pick the app your child or class uses most, and go down the list. You can answer most of these by using the app for ten minutes and getting a few answers wrong on purpose. The rest are questions for the company. If it can't answer them, that is an answer too.

The checklist

1. Evidence: what does it notice when my child answers?

Ask: When my child picks a wrong answer, does the app know which wrong idea that answer stands for?

A good answer: Each wrong option is built from a named mistake, such as adding instead of taking away, mixing up b and d, or taking the smaller digit from the bigger one. The app records which one was picked. It also notices answers too fast to have been read and treats them as guesses, not knowledge.

A weak answer: "We track right and wrong." A wrong answer that carries no information wastes a third of what three-option questions can tell you.

Peabody: Wrong boxes in maths facts, phonics and the upper maths rungs are named mistakes; number sense and sentence questions are not yet. Rapid taps, guess-and-check and option cycling count as no evidence.

2. Learner model: how does it decide what my child knows?

Ask: What decides that my child "has" a skill, and how sure is it?

A good answer: A statistical model, with the number of answers it needs and how it handles lucky guesses. The best ones keep an error bar, so one lucky tap doesn't move a child up. A 2026 benchmark found that large language models asked to do this job on their own barely beat "always guess right".

A weak answer: "Our AI understands your child." No product can show that.

Peabody: Ordinary code: item levels from 0 to 5 and a ladder of rungs per subject, with no error bar yet. Our knowledge tracing explainer compares the options.

3. Content: who checks the questions?

Ask: If the app writes questions with AI, what checks them before my child sees one?

A good answer: A second route that doesn't trust the writer: a solver for maths answers, a list of allowed words for early reading, a person or a source for facts, and a safety check.

A weak answer: "The model is very accurate."

Peabody: The practice questions are generated by code from fixed banks, not by a language model. The stories are written by AI and screened.

4. Choosing what's next: can it make things easier?

Ask: If my child starts getting things wrong, how quickly does it get easier, and can it go back down a level?

A good answer: It aims for most answers to be wins (about 80 to 85% on new material; see the 85% rule), steps down after a few misses, and changes approach for a child who is stuck instead of serving the same kind of question an eleventh time.

A weak answer: It only ever moves up, or a grown-up has to email support to reset a level.

Peabody: After ten tries without three right in a row, it asks the question another way, then goes down a rung, then shows a worked example. A game's own level can drop three steps but climbs only one above grade. A rung a child has passed does not yet re-open on its own.

5. Experience: is the learning part of the fun or in its way?

Ask: Can my child keep playing after a wrong answer?

A good answer: Yes. A question costs a little time and never ends the game, and the child can choose a mode with no questions at all.

A weak answer: The game stops until the question is right. That turns a hard question into a wall.

Peabody: Games ask rarely by default (the Light dial), a miss costs tempo, not the run, and "Just play" asks nothing.

6. Tutor moves: what happens right after a mistake?

Ask: After a wrong answer, does the app explain the mistake, or just say "try again", or just show the answer?

A good answer: One line that names the mistake, a hint that doesn't give the answer away, and one more try. The 2026 trials found this is the moment AI help pays off, and only when the app offers it unprompted: in the Khanmigo trial students opened the tutor in only 17% of sessions where they made a mistake (NBER).

A weak answer: A chat box the child has to open, or an AI that will give the full answer on request. In a PNAS study, students who practised with an answer-giving chatbot did up to 17% worse once it was taken away.

Peabody: After a miss, a line made in code (no AI) names the mistake and draws the model; the retry hint never says the answer.

7. Grown-ups: does it tell you one thing to do?

Ask: What will I learn about my child, and what should I do with it?

A good answer: One plain sentence about what's solid, what's shaky, and one thing to try tonight. In the Tutor CoPilot trial, advice to the adult helped weaker tutors most.

A weak answer: Charts of minutes and streaks.

Peabody: A free grown-up account shows "one thing to try tonight", built from the rung a child is on or the mistake their wrong boxes keep naming. See where is my child stuck?.

8. Measurement: how does the company know it works?

Ask: What is the outcome you measure, and was it measured without the AI there?

A good answer: A later question, answered alone, on something not practised that day. For outside evidence, an independent randomized trial with the whole enrolled group, not a "used it as recommended" slice. Our evidence tiers cheat sheet explains the levels.

A weak answer: Time spent, lessons completed, or a satisfaction survey.

Peabody: Spaced checks about 3 days, 2 weeks and 2 months after a question retires, and an internal page that reports how many come back right. No independent trial yet.

9. Safety: what does it ask for and keep?

Ask: What personal details does it collect, how long does it keep them, and does it use my child's data to train AI?

A good answer: As little as possible, a written retention limit, and no AI training on children's data without separate permission from a parent (the amended US children's privacy rule now requires that permission).

A weak answer: A long privacy policy with no plain answer.

Peabody: Children choose a made-up name; the practice records answers and timing, not voices or faces. Our AI policy lists what is sent where.

Scoring it

Count the good answers. Seven to nine: the app is built the way the evidence points. Four to six: useful practice, but check how it handles mistakes (question 6) and getting easier (question 4) before relying on it. Three or fewer: it may still be fun, but treat it as a game, not a tutor.

If you'd like to see what these layers feel like for a child, start free in any browser or pick a story from the library. Our blog post on what state-of-the-art AI learning software looks like walks through a whole session.

Questions parents ask

Is an AI tutor better than an adaptive practice app?

Not on its own. In a two-year trial in 18 middle schools, adding the Khanmigo tutor to Khan Academy practice gave about what the practice gave without it. In a 6,000-student study, AI help added most when it came right after a mistake inside mastery practice. The structure does the heavy lifting; the AI helps at the moment of error.

Should my child use a general chatbot for homework?

Only as a checker after they have tried on their own. The OECD's 2026 report recommends learning without AI first, then with an educational AI built to give hints, and only then with general chatbots. A chatbot that hands over the answer raises today's score and lowers next week's.

How can I test question 4 myself?

Get three or four answers wrong on purpose in a row and watch what happens next. A good app makes the next question easier or shows a worked example. A weak one repeats the same kind of question, or keeps the level where it was.

Does Peabody use AI to decide my child's level?

No. The level, the schedule and what comes next are ordinary code with fixed numbers. AI writes and screens the stories. That keeps the part that decides what a child sees predictable and checkable.

Free at peabody.kids/guides/ai-learning-software-checklist-nine-layers