LUMORA

SHORT EDITION · 2026

The Science Behind How We Learn

What the research actually shows about how humans learn — and what it means for AI tutors

LUMORA

An Apodeixis product

AN EVIDENCE REVIEW · 2026
COMPILED WITH AI ASSISTANCE

Contents

PART I · HOW LEARNING WORKS

01 Why this matters 5
02 What learning actually means 7
03 The illusion of learning 10
04 Retrieval 13
05 Spacing 16
06 Feedback 18
07 Understanding and visuals 21
08 Cognitive load and scaffolding 24
09 Metacognition 28

PART II · TUTORING AND MACHINES

10 Adaptivity and intelligent tutoring 32
11 One-to-one tutoring 36
12 AI tutors 40
13 The design brief 45

PART III · CLAIMS AND PROOF

14 Popular claims, audited 51
15 What would prove it works 54
16 Conclusion 58
A *How to read a claim* 61
B *Sources* 64
C *Lumora, feature by feature* 73

THE SHORT EDITION

A shorter book, same evidence

This is the short edition of *The Science Behind How We Learn*. The long edition runs to seventeen chapters and four appendices. This one keeps the argument and the numbers that carry it, and drops the literature-review scaffolding.

It assumes no research background. Every figure is real, every source used was checked against Crossref or a primary record, and where the evidence is thin the text says so rather than filling the gap.

The one-sentence version. The techniques that produce durable learning are well established, the feeling of learning is not a reliable guide to them, and the variable that decides whether an AI tutor helps or harms is whether the learner does the work.

Three findings to carry

Difficulty is the mechanism, not the failure. Students in an active-learning physics class learned more than students in a polished lecture and felt they had learned less (Deslauriers et al., 2019). The effort of retrieval is what encodes the memory; fluency is what prevents it.

AI tutoring helps performance, and can harm learning, for the same reason. A well-designed tutor beat active-learning instruction by 0.63 SD in a Harvard trial. An unguarded assistant left students 0.19 SD *below* peers who never had it on an unassisted exam (Bastani et al., 2025). The difference was not the model. It was whether the tool did the learner's thinking for them.

The composition is untested. Retrieval, spacing, worked examples, feedback and the rest are individually well supported. No study has tested whether one product's particular combination of them works. An intervention assembled from validated parts inherits the validation of the parts. It has not been validated.

If you have forty minutes

Read Chapter 3 (why learning feels like learning and usually isn't), Chapter 12 (what the AI tutoring trials actually found), and Chapter 16 (the conclusion). Those three carry the argument. The rest is the evidence.

Every research number is sourced. The full reference list is Appendix B.

PART I

How learning works

The mechanisms that produce durable learning — and the illusions that hide them.

CHAPTER 01

Why this matters

Most people have strong intuitions about learning, and most of them are wrong in ways that are invisible from the inside.

Reading a chapter twice makes it stick. Highlighting helps. An explanation that feels clear means you understand it. These beliefs are widely held, rarely examined, and occasionally expensive — because a great deal of money is spent on products built to satisfy them.

What the evidence actually establishes

Sixty years of learning science has produced a body of findings that replicate unusually well: retrieval practice, spacing, worked examples, the expertise reversal, feedback hierarchies, the multimedia principles. These are not contested in the way most of education is contested. They have large samples, increasing preregistration, and effects that survive replication.

What has not happened is anyone testing whether the whole set, applied together by one system, produces better learning than a well-chosen subset. This is ordinary in an applied field. We know a great deal about which interventions help individually, and rather little about whether a specific product's combination of them is the right combination.

The inference to resist. An intervention composed of well-supported parts is not thereby well supported. The parts may not be additive, may not transfer to a different population, and may interact — sometimes helpfully, sometimes not.



The measurement problem

Learning is not directly observable. We see a score, a time, a completion rate, a self-report, and infer a state of knowledge. Every inference is lossy, and the losses lean consistently in the direction that makes products look good.

Two consequences govern the rest of the book:

- **A measure taken immediately after instruction is systematically optimistic**, and the optimism grows as the delay lengthens.
- **A measure taken with the tool available is not a measure of what the tool taught.** It is a measure of what the learner could do with the tool.

Neither is a subtle statistical caveat. Both are large enough to reverse a conclusion, and both have reversed published ones.

What this book does not do

It does not market a product. Lumora is the tutor this review was written alongside, and Appendix C maps its features to the evidence without treating that map as validation. Where a technique is well supported, it says so. Where the composition is untested, it says that too. Where a popular claim is false, it says so — including when the claim is convenient for the field this book comes from.

The standard applies to Lumora's own work. A reader who turns the book's tests on the product that prompted it is doing the right thing, and Chapter 15 describes what that test would have to look like.

Chapter 2 does the unglamorous prerequisite work: defining what a learning intervention is supposed to change.

CHAPTER 02

What learning actually means

02

Before asking what works, it is worth being precise about what a learning intervention is supposed to change. The vagueness of the word “learning” is the single largest source of disagreement in this field, and most popular claims dissolve once the outcomes are pulled apart.

Six things that are not the same

OUTCOME	WHAT IT MEANS	HOW OFTEN MEASURED
Performance	Getting it right now, with the material in front of you	Constantly
Recall	Pulling content out of your head without support	Often, in research
Retention	Recall after a delay — the delay is the point	Rarely
Understanding	Explaining why, predicting under change, detecting falsehood	Poorly
Transfer	Using it somewhere new	Almost never
Metacognitive accuracy	Knowing how well you know	Almost never

Performance is cheap to produce and nearly useless as evidence of learning, because it can be carried entirely by cues present at the moment of test. A student who has read a worked solution can reproduce it. That reproduction is performance.

Understanding is what most people mean by “learning,” and it is the hardest thing to measure. There is no agreed instrument for it, which is not a small problem given how much is claimed about it. Transfer is the only outcome that matters for real life, and it is almost never measured.



Why the distinction is not pedantry

Two interventions look identical in a product demo.

A: the tutor teaches a concept, asks a four-option question, marks it, and congratulates the learner.

B: the tutor teaches a concept, asks the learner to type the definition from memory, marks it, and explains.

Both “check understanding.” The literature does not treat them as equivalent. Chapter 4 explains why the difference is not academic — and why the answer is more contested than either side admits.

> **The practical rule.** When someone says a product improves learning, ask which of the six things it improved, and how long afterwards. Most marketing copy cannot answer either question.

The forgetting curve is the floor

Memory does not hold still. Whatever is learned decays at a rate that depends on how much was learned, how it was learned, and how much has been retrieved since. “How much do I know” is therefore not a fixed property of a person. It is a function of time.

This sounds obvious and is routinely violated in software. A mastery score that rises with correct answers, falls with wrong ones, and never changes with time is exactly such a violation — and it is the most common way a learning product ends up lying to its user without intending to.

One example

You are learning the causes of the First World War for an exam in three weeks.

Performance is answering 12 of 15 multiple-choice questions correctly the night before, having read the chapter twice. You will probably manage this.

Retention is naming the seven alliance commitments on 1 March, three weeks later, with the book closed.

Understanding is explaining why the alliance system turned a local Balkan crisis into a world war, and what would have had to be different for it not to escalate.

Transfer is recognising the same logic — a defensive alliance turning a local conflict into a global one — in a crisis that was never in your textbook.

Metacognitive accuracy is knowing, on 1 March, which of those five you actually have.

Almost every method people use to study sits at the first one. Almost every method that works sits further right.

Chapter 3 is the one that matters most: why the things that feel most like learning are the ones least likely to produce it.

CHAPTER 03

The illusion of learning

This is the most important chapter in the book. Most disputes in education, and most design decisions in learning software, come down to it.

The finding

In 2019, Deslauriers, McCarty, Miller, Callaghan and Kestin ran a randomised experiment in large Harvard introductory physics courses. Same materials, same content, two conditions: active learning following best practice, or a traditional lecture by experienced, highly rated instructors. Students completed a pre-test, the lesson, a post-test, and a survey about how much they felt they had learned.

Students in the active classroom **learned more and felt they had learned less** (Deslauriers et al., 2019). The negative relationship was traced to the cognitive effort active learning demands: students read the difficulty of the work as evidence that learning was going badly, when the difficulty was the mechanism.

Two further findings matter. Students who perceived the instructor as *fluent* reported a stronger feeling of learning by more than half a standard deviation. And self-assessment got worse the *less* prior knowledge a student had: novices were using the smoothness of the instruction as evidence, and smoothness is not evidence.

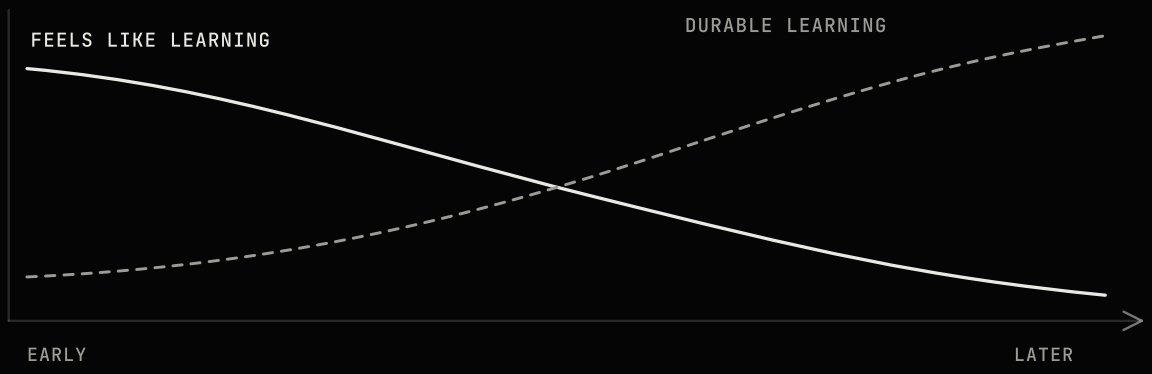
The calibration problem has a single number. In Karpicke and Blunt's (2011) experiment — the one where retrieval practice beat elaborative concept mapping on delayed science tests — **90 of 120 students (75%) misjudged their own learning** on that material. Three-quarters, on a task where the learning was real and substantial.

Why difficulty is misread as failure

Fluency and learning are different quantities, and the judgments learners make track the first.

A well-written explanation produces a specific feeling: things are resolving, effort is low, understanding seems to be arriving. That feeling is real. It corresponds almost perfectly to *processing fluency* — how easy the material is to take in right now. It is a poor guide to retention, because ease of processing leaves almost nothing to reconstruct later.

Struggling with a problem produces the opposite feeling. Effort is high, progress is slow, the work is unpleasant. That unpleasantness is, in a substantial number of well-controlled studies, a reasonable proxy for the eventual learning. The literature calls this a **desirable difficulty** (Bjork, Dunlosky & Kornell, 2013). The name is misleading as a slogan: what the research supports is narrower and more conditional than "harder is better." Chapter 8 returns to the conditions.



The four illusions

It helps to separate the phenomena collapsed under “feeling like you’re learning.”

Fluency illusion. Difficulty in the present is misread as incompetence rather than productive effort. The material felt hard; therefore I did not understand it; therefore I am not good at this. The last step converts a learning process into an identity judgment, and identity judgments reduce willingness to attempt hard things.

Illusion of competence. Fluency during study is misread as mastery. This is the rereading trap.

Illusion of explanatory depth. A smooth explanation feels like understanding because it is easy to process. Following an explanation and being able to produce one are different capacities, and the second is what teaches (Chapter 7).

Familiarity mistaken for mastery. Re-encountering material raises a feeling of knowing that is unrelated to whether you could produce it unaided.

FLUENCY effort read as failure	COMPETENCE smooth study read as mastery
EXPLANATORY DEPTH following read as producing	FAMILIARITY recognition read as recall

What the evidence says about study methods

Dunlosky, Rawson, Marsh, Nathan and Willingham (2013) reviewed ten common techniques. Practice testing and distributed practice earned high utility. Highlighting, rereading, summarization, the keyword mnemonic, and imagery generation earned low utility.

The low-utility methods are not useless. They are *actively counterproductive relative to their felt effect*. Rereading produces strong fluency; the fluency feels like mastery; the mastery feeling is why people keep rereading; and the fluency is exactly what prevents the durable encoding rereading was supposed to produce.

> **The trap in one line.** The methods that feel most like learning are the ones the evidence most consistently identifies as failing to produce it — and the feel is why they persist.

What this means for a product

Any measure a learner generates about their own learning is suspect. Confidence, satisfaction, time-on-task, perceived progress: all contaminated by fluency. This is why Chapter 15 builds its experiments around unaided delayed tests rather than user satisfaction.

A product that makes learning feel easy is at elevated risk of making it worse, and the risk rises as expertise falls — precisely the population a tutoring product targets. Bastani et al. (2025) found exactly this shape: students with unrestricted access to a capable model performed better with it and worse without it than students who never had it.

The design burden falls on the system. If users cannot reliably tell whether they are learning, a product that depends on them noticing will not work. The system has to build the difficulty in, the retrieval in, the delay in, and the unaided test in — because the user will not supply any of those and will actively prefer their absence.

That is a strong argument against the design most AI products ship, which optimises for the moment the user closes the tab.

Chapter 4 turns to the most reliable effect in the field — retrieval — and the conditions under which it works or quietly fails.

CHAPTER 04

Retrieval

04

Retrieval is the most reliable finding in the literature, and the conditions under which it fails are the most important engineering constraints on any tutoring system.

The landmark experiment

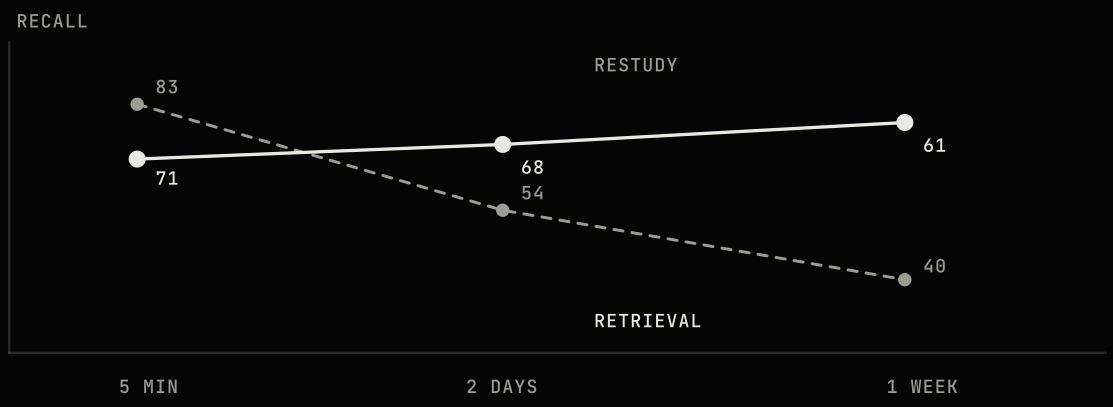
In 2006, Roediger and Karpicke gave students a prose passage under one of three schedules, all matched at four five-minute study periods: study four times (SSSS); study three times then take a recall test (SSST); or study once then take three recall tests (STTT).

The design is lopsided. Students in the repeated-study condition read the passage **14.2 times** on average. Students in the repeated-test condition read it **3.4 times** and spent the rest of the time struggling to recall it.

At five minutes, studying won: **83% for SSSS, 78% SSST, 71% STTT**. At one week the order inverted completely: **STTT 61%, SSST 56%, SSSS 40%**. The group that had read the passage less than a quarter as many times now recalled twenty-one percentage points more (Roediger & Karpicke, 2006).

The effect sizes are cleaner than the percentages: at five minutes, restudy won 81% to 75%, $d = 0.52$; at two days, testing won 68% to 54%, $d = 0.95$; at one week, 56% to 42%, $d = 0.83$. The Learning Condition × Retention Interval interaction was $F(2,117) = 32.10$, partial $\eta^2 = .35$. The crossover is the finding.

The uncomfortable implication. A system that checks understanding immediately after teaching, and updates a mastery number from that check, is measuring the side of the crossover where restudy wins. A check taken only at that moment cannot see the difference between durable learning and the feeling of it.



The general effect

Adesope, Trevisan and Sundararajan (2017) pooled 118 articles, 272 independent effect sizes and 15,427 participants. Practice testing against non-testing conditions: **g = 0.61** fixed-effects (95% CI [0.58, 0.65]), **g = 0.70** random-effects. Against restudy and rereading specifically: **g = 0.51**. Heterogeneity is high (I^2 = 0.81), so the average is solid and the effect in any given context is genuinely uncertain. Yang et al. (2021) replicated the effect in real classrooms, which matters more for a product than word-list studies do.

Feedback is not optional

Retrieval practice is materially less effective without feedback, particularly in multiple-choice format. A multiple-choice test with no feedback can **confirm a wrong answer**: select a lure, learn nothing, and leave believing you were right. In the format, an incorrect selection confirmed without feedback produces a documented misinformation effect, with intrusions of the lure on a one-week delayed test.

Feedback is more valuable after an incorrect response than a correct one. There is simply more to say about a miss.

Is multiple choice as good as free recall?

This decides how much of a product's reported progress can be trusted, because most software practice is multiple choice.

The honest answer is that the question is a moderation, not a verdict.

When the practice format is the moderator, multiple choice is not weaker: Adesope et al. found larger effects for multiple-choice practice (g = 0.70) than short-answer (0.48), free-recall (0.62) or cued-recall (0.58). **When the final-test format is the moderator**, the ordering reverses: effects were largest when the final test was free-recall (g = 0.71) and smallest when it was multiple choice (g = 0.56). Rowland (2014) reported the opposite pattern to Adesope on practice format, probably because he excluded applied classroom studies.

The cleanest evidence is transfer across formats, a true manipulation. Pan and Rickard (2018) found transfer of **d = 0.58** across test formats, the largest of their six categories. The critical moderator was not format but **response congruency**: d = 0.28 without it, 0.58 with it, and 0.78 with elaborated retrieval. With neither, 0.21.

Smith and Karpicke (2014) resolve most of the dispute. Across four experiments, multiple-choice retrieval produced retention and inference benefits relative to study-only, with little or no advantage for short-answer in **three of four** experiments. The short-answer advantage appeared only when initial retrieval success was low.

So: keep multiple choice for efficient checking where learners are succeeding, and require production precisely where the check shows they are not.

Transfer is the limit

Retrieval practice reliably improves retention of *what was retrieved*, and reliably fails to improve performance on *what was not*.

TRANSFER TYPE	EFFECT	SIGNIFICANT?
Test format	d = 0.58	yes
Mediator / related cues	d = 0.61	yes
Application and inference	d = 0.32	yes
Problem-solving	d = 0.29	no
Stimulus–response rearrangement	d = 0.22	no
Untested materials	**d = 0.16**	**no**

The three non-significant categories are exactly the far-transfer ones. Adesope et al. show the same asymmetry by outcome: retention g = 0.63 across 243 effects, transfer g = 0.53 across **eleven**. That transfer stratum is statistically fragile and should be treated as suggestive.



A tutoring system that generates questions from the same material it taught will produce excellent retention of that material and near-zero transfer. This is not a flaw in the technique; it is what the technique does. It is also the shape of every flashcard app and every quiz-based app.

The design rule. Retrieve different things from different materials, and occasionally require production rather than selection. Do not expect the retrieval of one fact to build the structure that transfers.

In Lumora. A due review opens a short recall check, not a reminder to reread. The queue ranks attempted concepts by low mastery first, then by how long it has been since the last review.

Chapter 5 covers spacing — the second high-utility technique, and the one most often confused with a reminder.

CHAPTER 05

Spacing

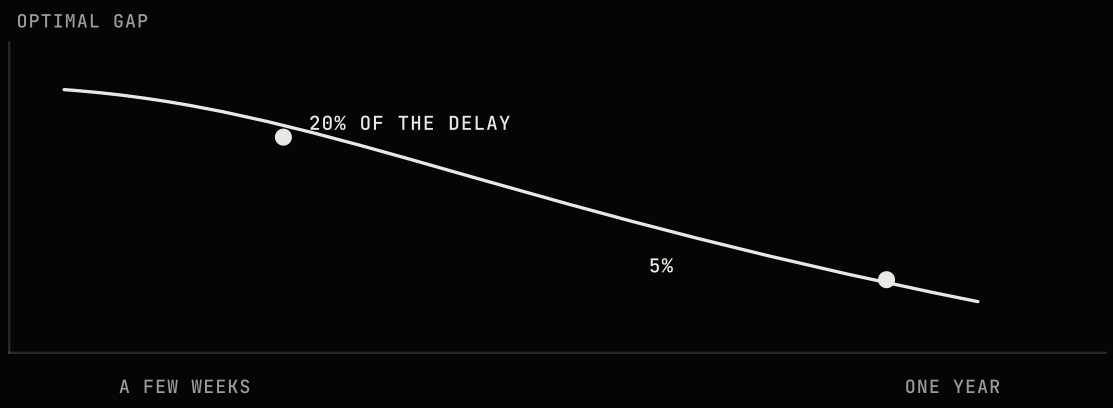
05

Spacing is the second of the two techniques Dunlosky et al. rated high utility. It is also the one most often confused with something else, and the confusion is where most “spaced repetition” products go wrong.

The basic finding

Cepeda, Pashler, Vul, Wixted and Rohrer (2006) synthesised several hundred verbal-recall studies. Everyone knows the headline: spreading practice beats massing it. The finding nobody quotes is more useful: **the optimal gap depends on how long you intend to keep the material.**

Cepeda, Vul, Rohrer, Wixted and Pashler (2008) established the shape with a large experiment — over 1,350 individuals, gaps up to 3.5 months, final tests up to a year later. At any retention interval, increasing the gap first improved and then reduced test performance. The optimal gap was approximately **20% of the test delay** for delays of a few weeks, falling to approximately **5%** when the delay was one year.



Two things follow, and both are frequently lost.

First, there is an optimum. More spacing is not monotonically better; space too much and the gap is too long to sustain through the intervening material.

Second, and counterintuitively, **longer retention horizons call for relatively tighter spacing**. The optimal gap is a *smaller fraction* of the test delay when you want to remember for a year. Intuition runs the other way, assuming long-term retention justifies long gaps.

Why a fixed schedule is wrong

A fixed “review every N days” policy cannot be right, because the optimal interval depends on the retention horizon, which depends on what the learner is trying to do — something the system rarely asks.

A learner who wants to remember something for a week and one who wants it for a year need materially different schedules. A single global setting cannot serve both. A system that asks for a target retention interval and derives its intervals from that is doing something more principled than one that picks a number.

A corrective on expanding schedules

Anki and most competitors use an expanding schedule: intervals grow as mastery is demonstrated. The folklore says this is optimal, with a roughly 10% advantage for expanding over equal spacing, originally derived from face–name learning.

That folklore does not survive contact with expository text. Karpicke and Roediger (2010) crossed the position of the first test with the spacing of repeated tests, using free recall on prose and a one-week delayed final test. Repeated testing beat a single test; testing with feedback beat testing without; **and there was no difference between expanding and equally spaced schedules**. Delaying the first test improved long-term retention.

For prose, the specific algorithm barely matters, as long as attempts are repeated and spaced. That is a useful negative result, and it should temper the confidence with which any product sells its scheduler.

What the evidence prescribes

Deliver the retrieval, not the reminder. This is the high-value change and it is cheap. When the system identifies a concept worth revisiting, it should open a session that asks the learner to retrieve it — ideally requiring production — and record the result. Nearly any spaced retrieval beats no retrieval.

Model forgetting. A per-concept, time-aware estimate of current retention replaces both a raw mastery number and a recency heuristic. Half-life-regression schedulers are established engineering, not research.

Ask for the retention horizon. It is one cheap question, and it makes the schedule defensible rather than arbitrary.

> **The distinction that matters.** A reminder to review is not spaced retrieval. A notification that navigates to a list of topics has placed the effort, the decision and the abandonment risk on the learner — the least reliable actor in the system for deciding what to study.

Chapter 6 covers feedback: the most double-edged intervention in the field, one that reliably works and that a tutoring system can easily make harmful.

CHAPTER 06

Feedback

06

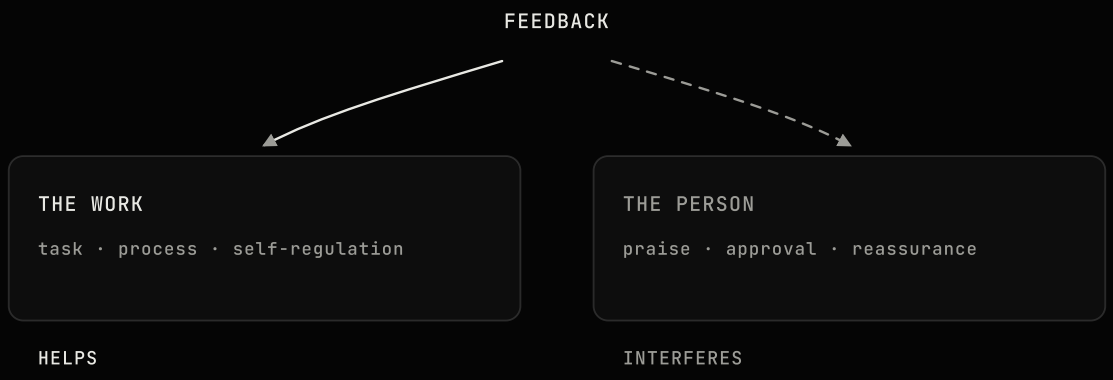
Feedback is the most reliably effective intervention in education and the easiest to deploy badly. A tutoring system that gives generic encouragement after every answer is not a harmless default. It is an intervention with a documented capacity to make learners worse.

What the evidence establishes

Kluger and DeNisi (1996) meta-analysed 607 effects across 23,663 observations. The mean effect is usually reported as **d = 0.41**, and the more consequential finding is the distribution: in more than a third of the feedback interventions reviewed, feedback **reduced** performance.

This is not a paradox. Their explanation is that feedback works by directing attention, and the direction determines the outcome. Feedback aimed at the *task* — what was right, what was wrong, what to try — helps substantially. Feedback aimed at the *self* — reassurance, praise, approval — directs attention inward, where there is nothing useful to work on, and interferes with the processing that would have produced learning.

The practical rule. Feedback that helps points at the work. Feedback that hurts points at the person.



Hattie and Timperley (2007) organised feedback into a hierarchy that remains the most useful framework in the literature: task level (“your answer is right; here is the reasoning”), process level (“you got the method right but didn’t check the sign”), self-regulation level (“here is how you could check your own work next time”), and self level (“good job”). Most software feedback occupies only the weakest level. Process and self-regulation levels are where the value lives.

Wisniewski, Zierer and Hattie (2020) re-analysed the literature at greater scale and refined the message: content and type matter most, and cognitive and motor-oriented feedback outperforms motivational or behavioural feedback.

The praise problem, stated precisely

Praise is a self-level message. By Kluger and DeNisi's mechanism it should be at best neutral and at worst interfering. What the literature supports is *informational* acknowledgement — "good, that is the standard approach" — which is about the work, not the person.

For a tutoring system the risk is specific: an LLM asked to be encouraging will produce a lot of praise, because that is what conversational systems are rewarded for sounding like. "Never praise-only" is worth stating explicitly in a system prompt.

Timing: immediate versus delayed

The folklore is overwhelmingly pro-immediate. The evidence is more mixed.

The standard justification is that immediate feedback corrects an error before it is encoded. The counter-argument has empirical support: an immediate correction is available at the moment of the error, so the learner can use it without ever having to retrieve the correct answer. Delay it and the learner must first attempt retrieval, which is the thing the whole Chapter 4 literature says does the work.

The resolution is a distinction between functions:

- **Feedback on an incorrect answer** should be immediate. The learner has a specific misconception and needs it named.
- **Feedback on a correct answer** can be delayed or omitted. The retrieval already happened and the answer was produced.

The sharpest caution: feedback can make things worse

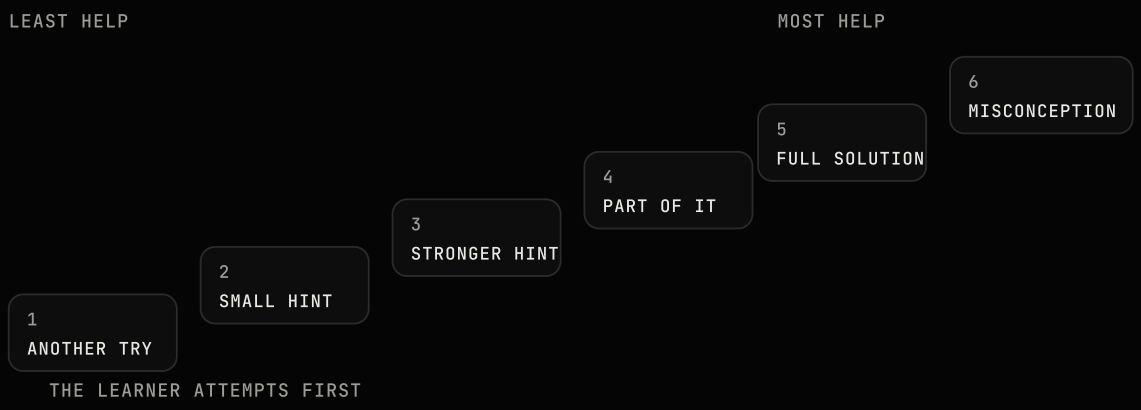
"Give feedback during problem solving" is treated as a safe default and it is not. Fyfe and Rittle-Johnson (2016) found that children with *some* prior knowledge learned **more without feedback than with it** during problem solving. A learner told the answer mid-problem has not first generated a solution, and the substituted answer enters with the same unwarranted confidence.

Explanations layered on top of a learner's own reasoning buy a small effect — on the order of $d = 0.16$ in Wittwer and Renkl's work — another argument for letting the learner produce before the tutor elaborates.

Multiple-choice sharpens the stakes. An incorrect selection confirmed without feedback produces the misinformation effect described in Chapter 4. Feedback is not a polish item in a multiple-choice tutor; it is load-bearing. A system that marks an answer without explaining it is running a mechanism known to implant the distractor.

The hierarchy the evidence implies

1. Encourage another attempt.
2. Give a small hint.
3. Give a stronger hint.
4. Show part of the solution.
5. Show the full solution.
6. Name and explain the misconception.



Steps 1–4 are graduated hints, well supported: the learner attempts before receiving, and help scales with demonstrated need. Step 6 is where process-level feedback lives, and it needs an explicit rule — “never restate the same words; teach the misconception behind them” — because a tutor that restates the answer has silently dropped from step 6 to step 5.

Record which items were answered wrongly, and reinject the *misconception* into later work rather than the question. Closed-loop misconception targeting is one of the better-evidenced design moves available.

The confidence trap

Hypercorrection research found that when people answer confidently and are then told they were wrong, their corrected answer is frequently also wrong — more often than after a low-confidence error. Confidence is not repaired by being told one is wrong; the learner substitutes the new information with the same unwarranted certainty.

An incorrect answer from a confident learner is the most dangerous moment in the interaction. A system with no confidence signal at the point of correction is missing a variable the literature identifies as predictive. Capture confidence per answer; on a high-confidence error, ask for the learner’s reasoning before revealing the correction.

Chapter 7 turns to understanding: the difference between following an explanation and being able to produce one.

CHAPTER 07

Understanding and visuals

07

The hardest thing in learning science to measure is also the thing everyone most cares about. This chapter is about the gap between following an explanation and being able to produce one — and about the visual channel, which is where that gap is most often widened by accident.

Following is not producing

Read a good explanation of why the sky is blue. You will understand it. You can follow the reasoning, find each step sensible, and summarise it in language that sounds like an explanation.

A week later, asked to explain why the sky is blue with no text in front of you, you may produce nothing — or something assembled at the time rather than stored. Nothing went wrong. The understanding was real when it existed. What did not happen is **consolidation**: the transformation of a comprehended explanation into a retrievable one.

Fiorella and Mayer's (2015, 2016) survey of generative strategies — summarizing, mapping, drawing, imagining, self-testing, self-explaining, teaching, enacting — found positive effects across reviewed studies, with self-explanation among the better supported. The mechanism is straightforward: producing an explanation requires retrieving the relevant knowledge and structuring it, which is more demanding and more durable than reading a structure handed to you.

Bisra et al. (2018) meta-analysed self-explanation and found positive effects on inference and transfer relative to no prompting. The instruction to explain also works as **pretraining**: explaining before instruction activates the learner's existing model, so the subsequent instruction has something to attach to.

The boundary condition matters. Self-explanation requires enough prior knowledge to explain *with*. A true beginner asked to explain a concept they have never met produces confabulation, not schema. This is why the worked-example literature puts explanation *after* a model, not before.



Visuals: the rules and the warning

Paivio's (1986) dual-coding account proposes two partly independent systems — verbal and nonverbal — linked by referential connections. Memory improves when material engages both and the connections are used. Mayer's cognitive theory of multimedia learning turned this into testable claims about limited capacity in each channel.

The consequences that matter:

- **Words and pictures together beat words alone**, when designed as a coordinated whole rather than as decoration.
- **On-screen text that duplicates narration can hurt.** This is the redundancy effect (Mayer & Moreno, 1999). Reading a caption while hearing the same sentence consumes the same limited capacity twice. A voice-first tutor that speaks an explanation and displays the same prose is running the redundancy effect on every turn.
- **Segmenting helps.** Rey et al. (2019) meta-analysed 57 studies and found reliable benefits for breaking complex material into manageable chunks under system-paced conditions.
- **Cueing helps modestly.** Cromley and Chen (2025) found self-explanation at g = 0.46, scaffolding at 0.38, and cueing at 0.24. The ordering is informative: asking the learner to do something beats pointing at something.



The seductive-details problem. Relevant and interesting material that does not support the learning objective *hurts* retention. A meta-analysis covering 58 papers and 7,521 participants found that engaging material which does not serve the objective actively harms retention, not merely fails to help. A capable language model's default output is maximally engaging and largely irrelevant.

This is the deepest tension in any product built around an appealing conversational surface. The same capability that makes a tutor pleasant makes it less effective per unit of engagement. Those are not two properties in tension; they are the same property viewed twice.

A system needs a reasoning rule, not a frequency rule: include a visual only when structure would clarify faster than prose — graphs, forces and vectors, process flows, molecules, trees and data flows, timelines and maps — and prohibit decoration. Then teach *through* the visual: point at a part, ask what happens and why, evaluate, continue. Integrate labels rather than placing them separately; never pair a diagram with an unrelated wall of text.

Pretesting: ask before teaching

One generative strategy is genuinely counterintuitive: ask the learner to attempt material before it has been taught.

Richland, Kornell and Kao (2009) established the effect: rendering test items before study improved later learning, and the retrieval attempts were mostly unsuccessful and helped anyway.

Two independent meta-analyses agree on the shape. St. Hilaire, Chan and Ahn (preregistered) report a **specific effect of g = 0.54** and a **general effect of g = 0.04**. King-Shepard et al. (2025) replicate it: **specific g = 0.66, general g = 0.01**. Pretesting helps a lot on the specific things you prequestioned, and essentially not at all on anything else. Pan and Sana (2021, n = 1,573) found pretesting matched or exceeded posttesting across formats and delays.

The rule that follows. Probe before teaching only where the learner has enough background to produce a plausible guess, and always give feedback after. Do not expect transfer to untested content; teach that content directly.

In Lumora. Diagrams are generated beside the explanation and followed by a check. The correctness of a generated diagram is not verified; Chapter 13 treats that as an engineering limit, not a solved problem.

The measurement gap

A recognition-format assessment cannot measure whether learners understand anything. Without a free-response instrument, every check is a multiple-choice question graded by matching a selected index against a stored key.

So the learner's *production* of an explanation is never assessed, only their selection among supplied options. Whether they can *explain* is never tested. Conceptual and relational understanding is never distinguished from factual recognition. The distinction between a robust mental model and memorised options is never observable.

An optional "I was thinking…" note is a partial exception. But unless it is stored, parsed, scored and used to update the learner model, it is decorative: a sophisticated, partially incorrect argument and "idk" are recorded identically as an incorrect answer.

The generative layer can be in the prompt while the assessment layer cannot see it. That is the central tension of any tutor built this way, and it is not fixable by better prompting.

Chapter 8 returns to working memory, worked examples, and the expertise reversal — the mechanism that makes help which suits a novice actively harm an expert.

CHAPTER 08

Cognitive load and scaffolding

08

Some of the best-evidenced principles in learning science are *conditional on the learner's level of expertise*. This chapter is about the mechanism by which a technique that reliably helps a beginner actively damages a more advanced learner — and why it is the most under-implemented safeguard in adaptive tutoring.

Working memory, and what it implies

Sweller's cognitive load theory starts from a modest empirical premise: working memory is severely limited — roughly four chunks, or a small number of interacting elements — while long-term memory is effectively unlimited and precise.

Instruction is therefore a design problem of managing limited capacity. The theory distinguishes **intrinsic load** (inherent to the material at a given level of expertise), **extraneous load** (imposed by how the instruction is delivered), and **germane load** (resources devoted to schema construction).

Intrinsic load is not a fixed property of a topic. It is a property of the topic *relative to the learner's prior knowledge*. The same integration problem is elementary to someone with the schema and insurmountable to someone without it. This is the most important idea in the theory for a tutor, and the one most often misapplied.

Germane load is contested; many researchers, including some of the theory's own authors, argue it is a category error. The actionable version is simpler: **eliminate extraneous load, and the capacity that frees up goes to schema construction automatically.** A system optimising for a "germane load score" is optimising for a construct that may not exist.

Element interactivity

Sweller, van Merriënboer and Paas (1998) introduced element interactivity as the formal measure of intrinsic load: the number of elements that must be processed simultaneously. Instruction should present elements in a sequence where each step stays within working-memory capacity.

This is the theoretical justification for "one concept per message" — a correct rule, and one a capable language model will violate constantly, because verbose explanation is its default register and it has no model of the learner's current load.

The worked-example effect

Sweller and Cooper (1985) established the finding that surprised the field: students learning to solve new kinds of problems learned **more** from studying worked solutions than from attempting equivalent problems independently.

Problem solving consumes working memory on searching for a method, leaving little for learning. A worked example reduces element interactivity during the solution phase by presenting the steps in an order that requires no search, freeing capacity for schema-building.



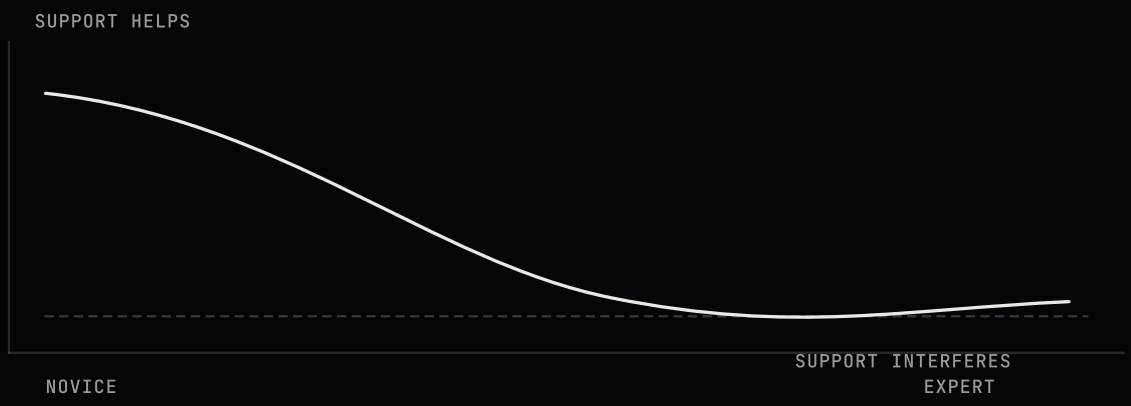
The instructional implication is **fading**: begin with a complete worked example, then progressively remove steps, then remove the example. The middle step — a **completion problem**, where the learner supplies the final step or steps — is where much of the learning happens. Jumping straight from worked example to independent problem removes support at exactly the moment the learner needs it most (Renkl & Atkinson, 2003).

Expertise reversal: when the help becomes the problem

Kalyuga, Ayres, Chandler and Sweller (2003) found that instructional support which benefited novices **harmed more knowledgeable learners**.

The mechanism is redundancy. For a novice, a worked example provides information they do not have. For an expert, the same example provides information they already have while consuming working memory and disrupting fluent processing. The explanation that was exactly right for a beginner becomes an obstacle.

Two corollaries. Scaffolding that is not removed becomes a dependency, not a support. And the optimal amount of guidance is a function of the learner's expertise right now, which means a static policy is wrong by construction.



This is why “fading” is a design requirement rather than a nicety. A system that always gives worked examples is making things worse for its more advanced users — exactly the people most likely to abandon it.

The assistance dilemma

Koedinger and Aleven (2007) named the trade-off. A tutor must decide, moment to moment, whether to help or let the learner struggle. Help too little and the learner fails and learns nothing. Help too much and the learner never practises the skill.

The correct resolution is neither extreme: provide help, but withdraw it before it becomes a crutch. The assistance dilemma is expertise reversal viewed from the tutor’s side — and it is the mechanism behind the AI tutoring findings in Chapter 12. The critical boundary in Bastani et al. (2025) is between a tutor that gives answers and one that does not. Koedinger and Aleven wrote the framework for that decision in 2007.

Guidance, not discovery

The instruction-versus-discovery debate is usually staged as a fight. The meta-analytic evidence resolves it in a way neither camp states.

Alfieri et al. (2011) separated discovery into two conditions and the sign of the effect flips:

- **Unassisted discovery: d = −0.38** across 580 comparisons — actively worse than the comparison conditions.
- **Enhanced, guided discovery: d = 0.30** across 360 comparisons — significantly better.

The distinction is guidance, not format. Students left to discover learn less than students given ordinary instruction; students given *guided* discovery learn more. This is expertise reversal seen from the other direction: support is necessary for novices and counterproductive for the already competent.

What the evidence requires

Novice or repeated failure → worked example, then faded completion, then independent problem. Advanced or consistent success → reduce scaffolding, increase difficulty. Never paste a diagram beside an unrelated wall of text. Fade guidance as mastery rises.

The weakness is that this is instruction, not behaviour. Nothing in a prompt-level system measures whether the tutor is at the right point on the sequence. “Repeated failure” has to be tracked in a learner model, and the mastery number such a decision would use is itself the thing Chapter 9 shows to be misleading.

In Lumora. A new concept opens with an explanation and a worked example before independent practice. Checks then move mastery; fading is not yet driven by a measured per-concept fluency signal.

Chapter 9 turns to metacognition: the calibration problem, and why a product that shows the learner a number has an obligation to be honest about what it means.

CHAPTER 09

Metacognition

09

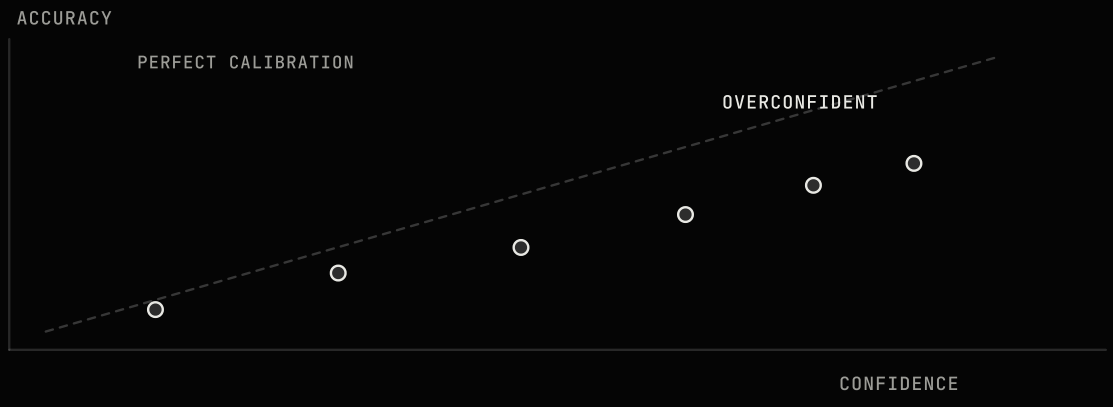
The most important thing a tutoring product shows its user is a number. This chapter is about what that number can honestly mean, and why the gap between the two is where a well-intentioned product misleads its user most easily.

What metacognition is

Flavell (1979) introduced metacognition as knowledge and cognitive monitoring about one's own cognitive processes. Two capacities are usually separated: **metacognitive knowledge** — accumulated understanding of one's strategies and their conditions of use — and **metacognitive monitoring** — the online judgment of one's own state. Calibration lives in the second.

Koriat's two-stage model describes retrieval as a fast stage that generates a feeling of accessibility, followed by a slower stage that retrieves the content. Learners report the *first stage* as confidence while the second determines accuracy. That is the mechanical reason confidence is a poor guide: it reports ease of access, not accuracy of memory.

The empirical direction is consistent. Learners systematically over-predict memory for material learned under easy, familiar conditions and under-predict for effortful, unfamiliar conditions. This is a robust bias, not noise, which means it predicts behaviour in the same direction most of the time.



The implication for a product. You cannot measure the value of instruction by asking learners whether it felt productive. A system that tunes itself to satisfaction is tuning toward the conditions the research says produce worse retention.

Prompts and the one causal win

If learners are bad at this, the obvious intervention is to prompt them. The evidence is more mixed than the prevalence of metacognitive prompting suggests.

One AI-tutoring evidence review found **metacognitive feedback was the one intervention with a clean causal win**: odds ratios of 0.47 for offloading and 1.51 for performance. In the same corpus, an **effort-based reward had no effect**. The effective intervention was not "tell the learner to reflect"; it was specific feedback about their own thinking.

The distinction matters. Many systems specify occasional confidence probes used to inform the learner model. That is the right shape, but two things are usually missing: the confidence data is not stored, and calibration is never computed. Calibration is a specific measurable quantity — the relationship between confidence and accuracy. A system collecting both could tell a learner something genuinely valuable and true: *you are systematically overconfident on recognition items and accurate on production ones.* Almost none do.

The illusion of competence

The illusion of competence is the error of treating fluent understanding as mastery. Three mechanisms reinforce each other.

The explanation is fluent. A well-written explanation is pleasant to read; the learner feels they understand. That fluency is the cue they misread.

The check confirms the feeling. The learner answers a multiple-choice question correctly and the interface confirms it. The test is in the same format as the study, so it inflates.

The number rises. The learner sees a percentage go up. A displayed score is where the system's own presentation amplifies the illusion, and the point at which a learner most reasonably concludes they know the material.

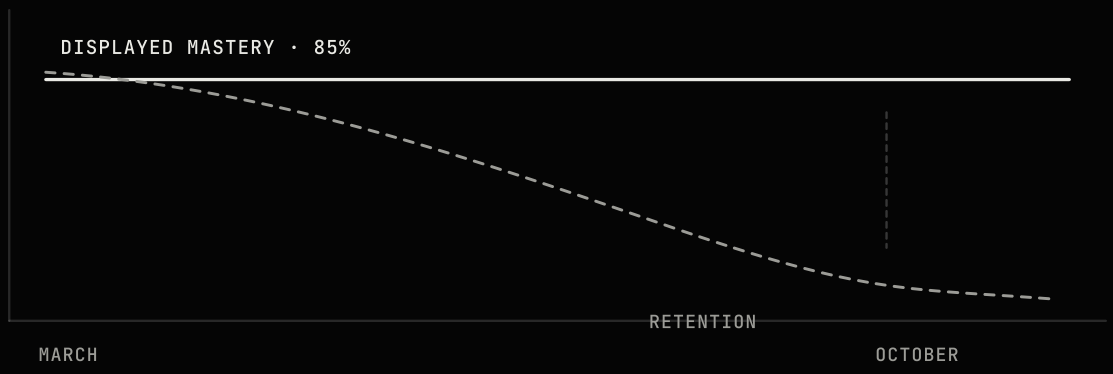
A system could break the chain at any point. Most do not.

Numbers that do not decay

A recurring technical failure is a performance counter that never ages. The typical shape is a number between 0 and 1 per concept. It moves up when the learner answers correctly, down when incorrectly, and is displayed as a percentage. It **never changes with the passage of time**. A concept demonstrated in March and untouched since still shows its March value in October.

The result is a display that reads as a fact about the learner ("you know this 85%") and is in fact a fact about a historical performance event ("you got this question right once"). Nothing in the interface distinguishes them.

There is a second problem. The number falls equally for two very different reasons a learner can be wrong: they do not know this, or they know a wrong model and are confidently applying it. A system that cannot distinguish a gap from a misconception cannot treat them differently, and the literature is unambiguous that they require different responses.



What a product owes its user

Distinguish what is measured from what is estimated. A mastery number derived from a handful of multiple-choice outcomes should not be presented as knowledge. If the system cannot estimate retention over time, show "last checked on 3 March, correct then" rather than a bare percentage.

Represent uncertainty. An adaptive tutor operates on an estimate and should behave as though the estimate might be wrong. Instructions that tell a tutor to teach confidently, without ever admitting uncertainty, instruct it to overstate what it knows.

Make calibration visible. Record confidence alongside correctness and, over time, show the learner where their sense of knowing and their actual knowing diverge. This is the highest-value unmet capability in the category, and unusually well supported.

In Lumora. Mastery is an attempt-derived estimate, and the review queue adds staleness. The number shown is not a retention model, and this book does not present it as one.

The meta-level finding. Humans are systematically wrong about which of their own study methods work, and the error is large, directional, and resistant to experience. A tutoring product is in the unusual position of being able to fix this at a scale no individual can manage. That is a stronger justification for building one than the usual pitch.

Chapter 10 asks whether personalization and adaptivity actually produce learning, or merely feel valuable.

PART II

Tutoring and machines

What a tutor does, what software can copy, and what the trials actually found.

CHAPTER 10

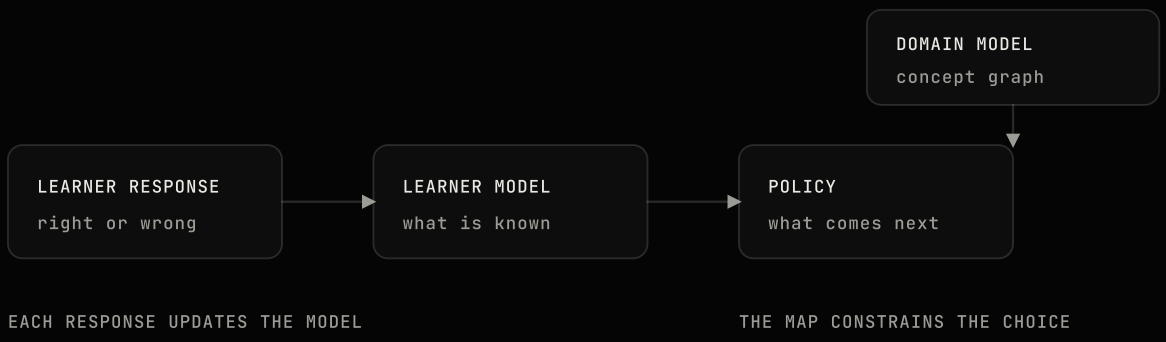
Adaptivity and intelligent tutoring

Does being adaptive actually produce learning, or is it a feature that feels valuable and measures approximately nothing? The literature's answer is more interesting than yes or no: adaptivity is real but modest, the strongest evidence is for *structured* adaptivity, and the popular claim is considerably stronger than any meta-analysis supports.

What an intelligent tutoring system is

An ITS is software that models a learner and adapts instruction to that model. The defining features are a **domain model** (structured representation of the material), a **learner model** (representation of what the learner knows), and a **pedagogical policy** (rules mapping learner state to instructional action).

A concept tree with prerequisites, a per-concept estimate, and rules mapping state to next action meet that bar in embryonic but real form. Expressing the policy as instructions to a language model is the novel and unvalidated part. It is still an ITS with a modern presentation layer.

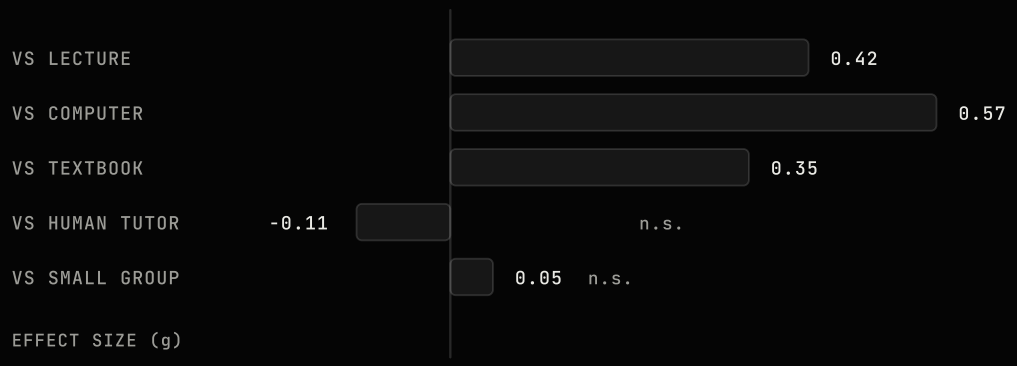


The meta-analytic verdict

Ma, Adesope, Nesbit and Liu (2014) synthesised 107 effect sizes from 14,321 participants:

COMPARISON	EFFECT
vs teacher-led large-group instruction	g = 0.42
vs non-ITS computer-based instruction	g = 0.57
vs textbooks or workbooks	g = 0.35
vs individualized human tutoring	g = −0.11 (n.s.)
vs small-group instruction	g = 0.05 (n.s.)

Against a lecture hall, an ITS is meaningfully better. Against a generic computer program, better still. Against a textbook, better but not dramatically — and a well-matched textbook is not a strawman. Against a **human tutor working one-to-one**, the ITS is not distinguishable from zero. The confidence interval includes effects in both directions.



VanLehn (2011) reaches a compatible conclusion: ITS effectiveness is roughly half that of human tutoring on average, with the gap narrowest in domains with clean procedural structure (algebra, programming, formal logic) and widest in ill-structured domains (writing, history, clinical reasoning). Kulik and Fletcher (2016) report a median gain of 0.66 SD across 50 evaluations — the most optimistic reading in the literature, and it depends heavily on whether the outcome is the tutor's own test or an independent standardised test. Effects on the tutor's own tests are larger. That is the same format-congruence inflation from Chapter 4.

The uncomfortable boundary conditions

K-12 mathematics ITS effects can be near zero. Steenbergen-Hu and Cooper's analysis of 34 independent samples reported effects commonly summarised in the g = 0.01–0.09 range. The college effects are considerably larger. Whatever adaptive software is doing well, it is not doing it in the setting where mastery curricula are most standard.

The K-12 result cuts against the equity argument. The same analysis found effects were *larger* for general-population students than for low achievers — the opposite of what a "close the achievement gap" case requires.

The pooled historical effect is decaying. Earlier studies reported significantly stronger effects than recent ones. A product shipping in 2026 inherits none of the historical pooled magnitude.

Does personalization itself help?

Sometimes, and less reliably than the market implies. Large-scale evaluations, including the RAND study of personalised learning environments, found that when implemented well it was associated with better outcomes, and that *implementation* mattered far more than the label.

The mechanisms with evidence behind them are specific:

- **Mastery-based progression** — advancing on demonstrated competence, not the calendar.
- **Prerequisite gating** — not presenting material before its prerequisites.
- **Adaptive difficulty** — task difficulty tracking demonstrated competence.
- **Formative assessment driving reteaching** — checks determining what comes next.

“Personalisation” as a label conceals which of these is actually present.

Knowledge tracing, and the data problem

Bayesian Knowledge Tracing models knowledge of each concept as a latent probability updated from correct/incorrect responses, with separate slip and guess parameters. Performance Factor Analysis extends it to partial knowledge; Deep Knowledge Tracing models the full interaction history. These models exist because estimating knowledge from correctness is hard, and a hand-tuned exponential smoother is not a solution.

Two deficits follow from the data, not the implementation:

It cannot distinguish knowledge from guessing. A correct answer on a four-option question has four possible causes, only one of which is knowledge. A boolean field collapses all four.

It cannot distinguish a lapse from a genuine gap. A learner who knew this and fumbled it is in a completely different state from one who never knew it, and the correct next action differs.

The diagnostic problem is hard even with proper tracing. Against 20,964 real student responses, automated misconception detection produced false alarms at roughly **eight to one** over genuine detections at realistic prevalence.

What adaptivity can and cannot be

Sorting the capabilities is more useful than arguing about the label.

Implemented as structure: a persisted concept tree with prerequisites; prerequisite gating in the interface; recommended-next selection; mastery and weak-concept injection into every session; recent-mistake injection as “address the misunderstanding behind this”; and instructions to binary-search the edge of understanding, treat all-correct as invalid evidence, and characterise a single wrong answer rather than react to it.

Prompt-level only, unverified: worked examples and fading; the feedback hierarchy; misconception-driven reteaching; confidence probes; difficulty adaptation tied to repeated failure.

Typically absent: a memory model with forgetting; a scheduler; knowledge tracing; uncertainty representation; any measure of whether adaptation improved outcomes.

The honest summary: **adaptivity specified as instruction is carefully reasoned and entirely unvalidated as a mechanism.** Whether the model complies, and whether compliance improves learning, are both unmeasured.

The distinction that matters most. The thing that predicts whether an AI tutor helps or harms is not how well it models the learner. It is whether it withholds the answer. An adaptive system that identifies that a learner needs a concept and then gives them the concept has been adaptive and has done nothing. One that identifies it and makes them struggle productively has done something.

Chapter 11 examines the one-to-one tutoring premise directly: what the two-sigma result actually shows, and which of its mechanisms are available to software.

CHAPTER 11

One-to-one tutoring

11

One-to-one instruction is often treated as the ideal form of teaching. The premise is nearly a century old and unusually well supported. This chapter takes it seriously, asks what actually makes it work, and asks which of those mechanisms are available to software.

Bloom’s two sigma problem

In 1984, Bloom compared ordinary classroom instruction, small-group instruction with a tutor, and one-to-one tutoring. The one-to-one condition produced gains roughly **two standard deviations** above the classroom average. In percentile terms, the average tutored student performed like the 98th-percentile student in a conventional classroom.

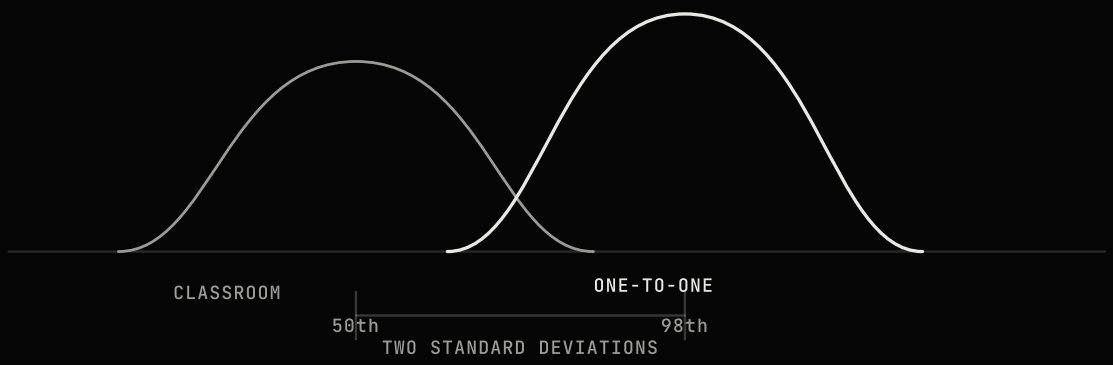
That figure is Bloom’s headline characterisation of a specific 1984 comparison, not a modern pooled effect size. The caveats matter.

Bloom’s tutors were exceptional. Graduate students trained for the task, working one-to-one over a short intensive period. The comparison is not one-to-one versus no tutoring; it is one-to-one versus the average classroom, which is a low bar.

The comparison group was ordinary instruction, not good instruction. The more demanding comparison is one-to-one against well-designed classroom instruction, and the gap narrows considerably.

The effect is real and replicated at scale. Low-cost structured tutoring programmes — notably “Teaching at the Right Level” work randomised in India — found substantial gains from targeted tutoring delivered by non-specialist volunteers. The mechanism does not depend on elite tutors.

But the large-scale wins came from *targeting learners by level and teaching at the right difficulty*, not from one-to-one-ness per se. Identification and level-matching were the active ingredients.



What actually makes one-to-one work

Diagnosis. A human tutor maintains a working model of what this student does and does not know, updated continuously. It is a model of this student's understanding, not a percentile.

Immediate, specific feedback. Not "good job" — "you have the right method but you dropped a negative, which matters when the quantity is negative."

Individualised pacing. The tutor waits. Most classroom instruction proceeds at the pace of the median student; a tutor proceeds at the pace of this one.

Prerequisite management. The tutor can shore up a foundation without derailing a class.

Scaffolding with judgment. The tutor reads whether this student needs the full worked example or can be asked for the last step, and decides per moment.

Accountability and attention. Someone is watching. The least studied and probably the most underrated mechanism, because a tutor's attention is a scarce resource no book or video provides.

Time on task. Bloom's tutors were simply present and working for long periods. Much of the gain may be sheer engaged effort.

Dialogue. The tutor asks questions and waits for answers, creating the generation Chapter 7 identified as central.

Which of these can software do?

Better than humans: consistency, patience, availability, and a complete record. A human tutor forgets what a student got wrong three weeks ago. A system with a persistent learner model does not. This is a genuine structural advantage and the strongest argument for the category.

Comparable to humans if done well: diagnosis from observed responses, pacing, prerequisite management, scaffolding decisions, feedback specificity, dialogue.

Worse than humans, and not obviously fixable: reading a student. A human sees confusion, hesitation, the beginnings of a wrong model forming in real time. A tutor that only sees correct/incorrect on four options is blind to nearly all of it. In one review, all 18 LLMs benchmarked on diagnosing cognitive skills from written work scored **F1 below 0.5**, and hallucinated evidence that was not present.

Much worse, or possibly negative: the risk of a confident, fluent, always-available answer. A human tutor does not have a plausible-sounding wrong answer to every question. One study found tutoring-mode dialogues entirely correct only **56.6%** of the time, against 85.5% for the same models' answers.

The assistance dilemma, restated

A human tutor resolves the help-or-struggle decision with social intelligence, contextual reading, and a long relationship. A software tutor resolves it with a policy. Every AI tutoring system has effectively made the same choice, and Chapter 12 shows that choice determines the outcome.

> **The paradox.** The thing that makes human one-to-one tutoring effective — the tutor's judgment about when to withhold — is the thing software is worst at, and the thing software is worst at is the thing the evidence says decides the outcome.

Bloom's problem restated

Bloom's tutors were not available, and that is the entire problem. The reproduction literature made progress by narrowing what the tutor does: not a brilliant general tutor, but one that targets a specific level with specific materials.

That is a design pattern worth taking seriously, and it is *not* the pattern an AI tutor naturally follows. Level-targeted instruction works by identifying a learner's level and delivering material pitched precisely at it, with uniform quality. An AI tutor works by being maximally flexible and conversational.

The design question survives the caveat: **would a conversational AI tutor be more effective if it were less flexible and more level-targeted?** The evidence offers no answer, and that is the honest position.

Flexibility is not the mechanism. Level-targeting is.

The Bloom gap, honestly stated

The best available ITS estimate against individualised human tutoring is g = −0.11: indistinguishable from zero. VanLehn puts ITS at roughly half the gain of human tutoring.

And yet the Harvard physics trial found a well-engineered AI tutor beating active-learning classroom instruction by a large margin. These are not in conflict. The AI tutor beat a good classroom and is still roughly level with a good human tutor. Bloom's tutors were better than anything currently fielded in software.

- Software tutoring beats ordinary and even well-designed classroom instruction.
- Software tutoring has not been shown to beat excellent human one-to-one tutoring.

- The gap is not primarily capability. It is judgment about when to withhold, and the risk of a fluent wrong answer.
- Closing it depends less on making the tutor smarter than on making it more disciplined about not answering.

Chapter 12 is the centre of the book: what the AI tutoring trials actually found, and why the positive and negative results are the same finding.

CHAPTER 12

AI tutors

12

This is the newest, least settled, and most consequential evidence in the book. It is also where a review has to be most careful, because the field is producing large effects in both directions, and combining them into a single optimistic or pessimistic number is exactly what this book exists to avoid.

The two landmark trials

The positive result. Kestin, Miller, Klales, Milbourne and Ponti (2025) ran a randomised controlled trial in an undergraduate physics course at Harvard, **N = 194**, comparing a custom AI tutor against an active-learning classroom with identical material. The tutor was GPT-4-based, with a content-rich system prompt encoding the same pedagogical practices as the in-class lessons.

The AI group's median post-test score was **4.5 against 3.5**, on a baseline of 2.75. The regression effect size was **0.63**, with quantile regression giving **0.73 to 1.3 SD**; the authors describe the linear estimate as an underestimate due to a ceiling effect. $p < 10^{-8}$. AI students also felt more engaged and motivated.

The negative result. Bastani, Bastani, Sungu, Ge, Kabakcı and Mariman (2025) ran a cluster field trial in high-school mathematics in Turkey with roughly **1,000 students** across three grades. Tutors covered about 15% of the curriculum. Three arms: a standard ChatGPT-style interface ("GPT Base"), a guardrailed GPT-4 ("GPT Tutor"), and a control.

- **With AI available**, performance improved sharply: **+48%** (GPT Base) and **+127%** (GPT Tutor).
- **On an unassisted exam after access was withdrawn**, GPT Base students scored **17% worse than controls who never had access at all** — approximately **−0.19 SD**. The guardrailed arm's harm was largely mitigated, at approximately **−0.01 SD**.

The authors' own mechanism statement: students used GPT-4 "as a 'crutch' during practice problem sessions, and when successful, perform worse on their own."



These are the same finding

Read together, the contradiction resolves. Bastani's assisted numbers are *larger* than the Harvard effect. Bastani's unassisted number is negative. The difference is not noise; it is the difference between measuring performance while the crutch is available and measuring learning after it is removed.

Deslauriers et al. established the general principle: students who learn more feel they learned less. Bastani et al. establish it in the form a tutoring product should fear — **students who perform better while the tool is present perform worse when it is not.**

The gap is quantifiable. A synthesis pooling comparable RCTs on **unassisted** outcomes reports a random-effects grand mean of **0.18 SD**, while with-AI and self-report measures run **0.50–0.67 SD**. That is the largest measurement artefact in this literature, and it is the gap between "the tool works" and "the tool taught anything."

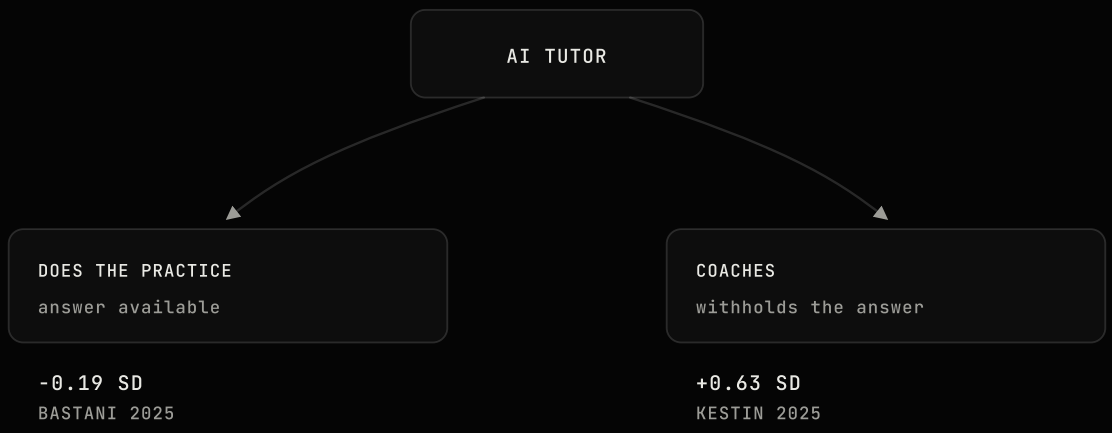
The design variable that explains everything

The most useful single document is a cross-trial synthesis of thirteen comparable randomised studies (Contractor & Reyes, IZA Discussion Paper 18792 — a working paper, not peer-reviewed, and it should be cited as such).

Its finding is not just the 0.18 SD. It is the explanation it attaches to the variance:

> Negative estimates come from designs where **AI could do the practice in the learner's place**. Positive estimates come from designs that **cast AI as a coach**.

That is the field in one sentence, and it is a design variable, not a capability variable. It explains why Bastani's unguarded arm is negative and its guarded arm neutral; why the Harvard tutor is positive; and why Hou et al. (2026) found that a five-step prompted interaction raised open-ended explanation scores by **2–3 points on an 18-point scale** ($p < 0.01$) while leaving multiple-choice recall unchanged — structure changed reasoning, not answer-getting.



The trials that look most like Lumora

Most of the discouraging evidence comes from unrestricted access to a general chatbot. The studies most relevant to a designed tutor share at least one of Lumora's constraints: the model is grounded in the course material, the interaction is structured, or the learner is asked to explain rather than copy. They are a different treatment, and they deserve to be read separately.

DESIGN	TRIAL	UNASSISTED RESULT
Course-grounded RAG tutor	Fischer, Rau & Rilke (2025), economics, N = 334	**+0.23 SD** overall; unrestricted access **+0.34 SD**, restricted access not significant. Working paper; one 25-minute session; immediate test.
Structured prompting around the same RAG tutor	Hou, Xiao, Liu & Mueller (2026), construction engineering, N = 95	Prompted group **2–3 points higher on open-ended explanation** (of 18, $p < 0.01$); no multiple-choice difference; unprompted ≈ slides. Preprint; small N; one course.
Metacognitive feedback on answer-offloading	Maier, Schwabe, Schneider & Feuerriegel (2026), fraction arithmetic, N = 704	Offloading **OR = 0.47**; unaided test **OR = 1.51**; effort reward null. Preprint; one-sided tests; the test-performance confidence interval reaches 1.0.
Platform tutor vs business-as-usual revision	Harrison et al. (2026), GCSE science, N = 929/644	**g = 0.33** (95% CI 0.18–0.48). Preprint; 30.7% attrition; immediate.
AI-assisted coding skill formation	Shen & Tamkin (2026), N = 52 developers	**50% vs 67%**, d = 0.738; largest gap on debugging; no significant speed gain. Company preprint; mostly junior developers; one library.

Two things follow. First, "AI tutor" is not one treatment. The designs that help are grounded, structured, and often supervised; the designs that hurt are unrestricted answer access. Second, structure is not the same as restriction. Fischer et al. found that continuous access to a course-grounded tutor beat a read-first-then-AI format; the unrestricted students integrated help gradually instead of bingeing it. Hou et al. found that prompting improved explanation scores but not multiple-choice recall. Maier et al. found that telling learners what offloading costs them changed behaviour without taking the answer away. For Lumora, the useful lesson is that the guardrail belongs in the interaction, not only in a locked door.

The full picture, including the unflattering results

Positive: the Harvard RCT (0.63–1.3 SD, immediate post-test only). Google's LearnLM trial in UK secondary schools found students at least as effective as those with human tutors, and better on novel problems (66.2% vs 60.7%), with delayed measurement — though it is a preprint and human-supervised. Stanford's Tutor CoPilot found students of AI-assisted human tutors were 4 percentage points more likely to master topics, and 9 points for lower-rated tutors — also a preprint, and about augmenting humans rather than replacing them.

Null: a Chinese grade-10 physics trial (N = 121 and 266) found effects on the unassisted end-term exam small and insignificant in all three AI arms (0.03–0.21 SD). A scaffolding study (N = 175) found no significant differences between full, light and no scaffolding; all arms gained. A marketing-course RCT found no significant differences in interest, self-efficacy or learning.

Negative: Bastani et al. (−0.19 SD unaided, unguarded). Abdelghani et al. (2026, work-in-progress preprint): post-test performance significantly *lower* than pre-test, N = 98 adolescents, with interactions dominated by instrumental requests and almost no monitoring or evaluation. Hou et al. (2026) found unprompted AI comparable to slides, with the prompted arm ahead only on open-ended explanation.

Preference diverging from learning: Yang et al. (2026, preprint) found voice interaction nearly doubled conversation length and cost 2.8× while producing mastery statistically equivalent to text — and students came to prefer voice. Engagement doubled; learning did not move.

Automation bias: Pitts et al. (2026, N = 432) found higher trust in the AI tutor predicted *lower* appropriate reliance. Trust is not a proxy for competence.

Performance and learning coming apart: Kazemitabaar et al. (2023, N = 69) found 1.67 SD on Codex-assisted training tasks, **no** detectable difference on the immediate unassisted post-test (0.05 SD), and 0.41 SD at one-week retention.

What no trial measured

No trial measured transfer to a novel domain. Not one. Every positive result is on the material that was taught, tested on the material that was taught or closely related material. A novel problem inside the same subject is still that subject.

Combined with Chapter 4's finding that transfer to untested materials is non-significant (d = 0.16) and the pretesting general effect of g = 0.01–0.04, the picture is consistent: the AI tutoring literature has established effects on **performance and retention of taught material**, and has produced no evidence about far transfer.

Delayed retention is rare. Kestin et al. measured an immediate post-test on the same material. The Bastani withdrawal test and the Kazemitabaar one-week test are the exceptions, and the one that measured delay properly is the one with the negative finding.

The factuality problem

Self-correction by a prompted model does not work. Kamoi et al. (TACL 2024) surveyed the literature and concluded that no prior work demonstrates successful self-correction with feedback from prompted LLMs, except in tasks exceptionally suited to it. A second call instructed to review the system's own claims and respond "OK" or "CORRECTION," with no ground-truth oracle, is prompted self-review in the configuration that fails.

Role labelling is a real lever, in an unexpected direction. Chen et al. (2026, preprint) found a model's willingness to correct an error is substantially gated by the chat-template role carrying the claim: relabelling a claim from an internal `<thought>` position to an external role raised explicit-correction rates by **23 to 93 percentage points**, significant in 10 of 12 settings. That supports passing a claim through a separate, externally framed call — but a check that runs on a plan rather than the delivered explanation leaves the explanation unchecked.

Tutoring-mode output is less reliable than answer-mode. One evaluation found LLMs correct on **85.5%** of college-algebra problems as answers, but only **56.6%** of their tutor dialogues entirely correct. A peer-reviewed study found ChatGPT-generated math hints failed quality checks on **32%** of problems. Generating an answer is a different task from sustaining a tutoring dialogue, and the second is harder.

What this means for anyone building one

Design for the crutch. Every discouraging result traces to unrestricted answer access. Every encouraging one traces to grounding, supervision, or withholding.

Both landmark trials used GPT-4-era models. The current Lumora build uses DeepSeek V4 Flash 0731 (V4F 0731), with the guardrail system described in Chapter 13 around it. A newer model and stronger guardrails are reasons to expect a different result; they are not evidence that one has occurred. The composition remains untested.

A system that always delivers a complete explanation is, by construction, the unguarded condition that produced −0.19 SD. A plan-then-teach pipeline that walks its plan node by node — explain, give an example, show a diagram, check — is a well-designed tutor. It is also, structurally, the unguarded condition.

This does not mean such a system will harm learning. It means the one study that tested the unguarded condition found harm, and a system resembling it has a specific, testable risk profile — which Chapter 15 turns into an experiment. The graduated hint hierarchy, the pretest gate, and “never turn every turn into an exam” are guardrails, and guardrails are exactly what separated GPT Tutor from GPT Base. The question is whether they are strong enough.

The honest summary

AI tutoring improves performance on the outcomes the trials measured. That is established. The effect size on immediate performance against weak comparators is large and sometimes dramatic.

But the variable that determines whether it produces *learning* is not intelligence, adaptivity, or conversational quality. It is whether the learner does the work. Every positive result in the corpus involves a system that structured the learner into doing the work. Every negative result involves one that did not.

And the corpus has produced **no evidence at all about far transfer**.

The honest public claim is narrow: *carefully engineered AI tutors can beat well-designed classroom instruction on immediate assessment of the material taught, and whether that becomes durable learning depends on whether the design makes the learner do the thinking.*

Chapter 13 sets out what an evidence-informed AI tutor should look like.

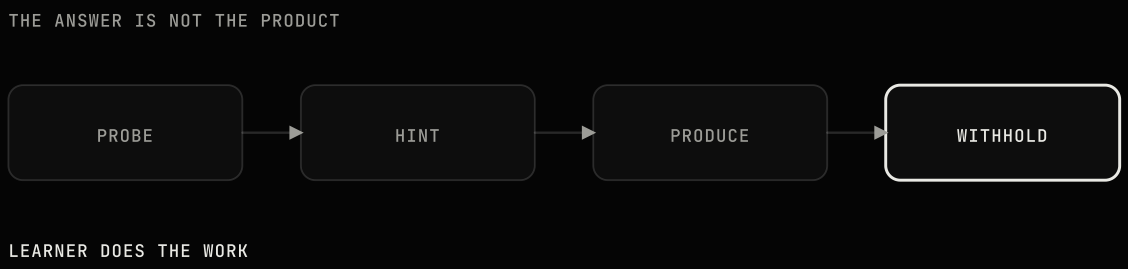
CHAPTER 13

The design brief

13

This is not a product pitch. It is a specification derived from the preceding chapters, and every requirement is tied to the study that requires it. Where the literature does not answer a question, it says **evidence insufficient** rather than inventing a number.

The organising principle comes from the AI tutoring evidence: what determines whether a tutor helps or harms is not how well it models the learner. It is whether the learner does the work.



1. Probe only where the learner can answer

Pretesting's benefit is almost entirely item-bound: **g = 0.54** specific and **g = 0.04** general in a preregistered meta-analysis, replicated at 0.66 and 0.01 (Chapter 7). The pretest gate must be a behavioural rule evaluated per question, not a global setting. A system that pretests everything is not pretesting; it is guessing, and the guessing is worse than the teaching it displaces.

Treat a run of all-correct answers as invalid evidence and escalate sharply. A run establishes a floor, not a ceiling.

2. Require production when retrieval fails

Response congruency, not format, is the dominant moderator: transfer is **d = 0.28** without it and **0.58** with it; 0.78 with elaborated retrieval, 0.21 with neither (Chapter 4). Format-congruent testing inflates measured learning.

Keep multiple choice for efficient checking where learners are succeeding. Require typed production — definition recall, explanation, "state the relationship between X and Y" — precisely when the check shows non-mastery. Grade transparently, showing what was matched so a learner can detect a wrong grade. Mix formats so the learner does not learn the test instead of the material.

Evidence insufficient: the optimal proportion of production questions.

3. Deliver retrieval, not reminders

A spaced-review system must open a retrieval episode, not send a notification. The distinction is not semantic. A prompt to remember places the effort, the decision and the abandonment risk on the learner — the least reliable actor in the system for deciding what to study.

Model forgetting per concept, with a stability estimate updated on each retrieval and intervals derived from a target retention probability. Ask for the retention horizon once. The optimal gap is about **20% of the delay** for a few weeks and about **5%** at one year (Cepeda et al., 2008, N > 1,350), so a single global interval cannot be right.

Do not assume the specific algorithm matters more than the spacing. Expanding and equal schedules were equivalent for prose (Karpicke & Roediger, 2010). Getting the episodes right matters more than optimising the algorithm.

4. Decide help versus struggle explicitly

Expertise reversal means support that helps a novice harms an expert. The assistance dilemma is the tutor-side view. The AI evidence shows the guardrail variable determining outcomes: −0.19 SD unguarded versus −0.01 SD guardrailed in Bastani et al.; and in Hou et al., prompted interaction improved open-ended explanation but not multiple-choice recall.

A graduated hint hierarchy should apply deterministically on repeated failure rather than at the model's discretion. Fade on **consecutive successful independent attempts per concept**, not on a composite mastery scalar: expertise reversal is about domain knowledge in the rich sense, and a multiple-choice mastery score is not a procedural-fluency signal. Escalate format and difficulty together when retrieval is failing.

Evidence insufficient: the optimal help/withhold threshold.

5. Sequence worked examples before retrieval for beginners

The least uncertain part of the specification. For a novice on a new problem type: worked example → completion problem → independent problem. For an expert: fade the guidance. Retrieval requires something to retrieve; a learner with no schema cannot retrieve from nothing.

The discovery evidence makes the same point from the other side: unassisted discovery is **d = −0.38**, guided discovery **d = 0.30** (Alfieri et al., 2011). The sign flips with guidance, not format.

6. Use a visual only when it carries structure

Words and pictures beat words alone when coordinated. Segmenting helps. Seductive details hurt across 58 papers and 7,521 participants. Self-explanation over visual material is g = 0.46, scaffolding 0.38, cueing 0.24.

Teach *through* the visual with a question, not past it. A visual the learner is told about is decoration; a visual the learner reasons from is instruction. Integrate labels into the diagram rather than placing them separately.

Check whether speech and text are complementary or duplicative. A voice-first system that speaks a complete explanation while displaying the same prose is running the redundancy effect on every turn.

Acknowledge that generated-diagram correctness is unmeasured. A structural audit catches malformed output; it cannot catch an inverted relationship. Constraining diagrams to forms whose correctness can be checked structurally — monotonicity, axis labels, node counts against a concept tree, edge direction against prerequisites — is engineering judgement, not evidence.

7. Aim feedback at the task, never the self

Mean feedback effect is d = 0.41, but more than a third of interventions reduced performance. Task, process and self-regulation levels help; self-level praise is the null and occasionally negative case. Never praise-only.

Feedback is not a safe default during problem solving. Children with some prior knowledge learned *more* without feedback than with it (Fyfe & Rittle-Johnson, 2016). Reserve feedback for the incorrect answer and the diagnosed misconception; do not volunteer it while a solution is being constructed.

Name the misconception, not just the correct answer. And capture confidence: when someone answers confidently and wrongly, the correction is frequently also wrong. On a high-confidence error, ask for their reasoning before revealing the correction.

8. Verify the text the learner reads

Tutoring-mode correctness is 56.6% against 85.5% for answer-mode on the same models; 32% of generated math hints failed quality checks. The explanations a learner reads are where hallucination risk is highest.

Verify against a retrieved source, not the model's own memory. Prompted self-correction without an oracle largely fails (Kamoi et al., TACL 2024). A separate, externally framed call helps — role relabelling moves correction rates by 23–93 points — but its reliability is a measurable quantity and should be measured against human ground truth.

The cheapest experiment in the book: run the same claims through the check, have humans judge which were genuinely wrong, and report the correction rate. It needs annotators, not participants, and it does not appear to exist in the literature.

9. Do not display confidence the system does not have

A number that rises with correct answers, falls with wrong ones, and never decays is a fact about a past performance event presented as a fact about the present. The learner cannot tell which they are reading.

In preference order: show a retention estimate with a confidence interval from a memory model; show "last checked on [date], correct then"; if a percentage is shown, make it decay and label it a performance estimate. Never imply certainty about the learner's state that the system does not have.

10. Record confidence and show calibration

Metacognitive feedback is the one AI-delivered intervention in the corpus with a clean causal benefit: odds ratios of 0.47 for offloading and 1.51 for performance. Effort-based rewards had no effect.

Capture a confidence rating with every check. After a few hundred checks, compute calibration and show the learner where their sense of knowing diverges from their actual knowing. A tutor that records both can say something true and useful: *you are overconfident on recognition items and accurate on production ones.*

Do not tune the system to learner satisfaction. Students who learned more felt they learned less; students preferred a voice modality that produced no learning benefit at 2.8× the cost. Satisfaction is not a weak proxy for learning here. It is an inverted one.

11. Test transfer separately from retention

Transfer to untested materials is d = 0.16 (n.s.); problem-solving transfer 0.29 (n.s.); pretesting's general effect 0.01–0.04. And no trial in the AI tutoring corpus measured transfer to a novel domain.

Every subject should include at least one assessment requiring application to a context never taught, reported separately from retention. **Evidence insufficient:** what transfer level an AI tutor can achieve. Any figure would be fabrication.

12. What a tutor should never do

- **Never give the answer the learner is practising retrieving.** The most robust predictor of harm in the literature.
- **Never present a performance estimate as knowledge.**
- **Never diagnose a misconception it cannot distinguish from a gap or a slip** — and say so when it cannot.
- **Never optimise for the learner's immediate satisfaction.**
- **Never claim an effect size it has not measured.**
- **Never say "science proves this works."** The techniques are individually supported; the composition is not, and no system has been tested.

In Lumora. Appendix C maps each requirement here to the feature that implements it — and marks the places where the evidence stops.

The specification in one paragraph

An evidence-informed AI tutor probes only where the learner can produce an answer; requires production as well as recognition, escalating format when retrieval fails; delivers spaced retrieval rather than reminders, scheduled by a model that forgets; gives worked examples to beginners and fades them for experts; aims feedback at the task and withholds it while a solution is being constructed; checks the explanations the learner actually reads, against a retrieved source; shows a retention estimate with uncertainty rather than a confidence-inspiring percentage; captures confidence so it can show the learner their own calibration; tests transfer separately; and does none of this while tuning itself to how the session felt.

That is buildable. Most of it is engineering. The parts that are genuinely research — whether the composition works, and how much transfer it produces — are what Chapter 15 is for.

Chapter 14 audits the popular claims in this field, because a product built on learning science inherits every myth the field repeats.

PART III

Claims and proof

How to audit a claim, test a tutor, and say only what the evidence supports.

CHAPTER 14

Popular claims, audited

14

Most claims about learning that circulate publicly are untested, overstated, or true in a weaker form than the claim states. This chapter audits the ones that matter, because a book that applies rigorous standards to the techniques it recommends while taking “learning styles” seriously would be incoherent.

The recurring methodological villain is **measurement substitution**: measuring what is easy — performance, preference, assisted output, self-report — in place of what matters — retention, transfer, unassisted learning.

The scorecard

CLAIM	VERDICT	NOTE
Learning styles / VAK	**False**	No reliable matching effect
10,000 hours	**Strongly overstated**	~14% of variance; ~5% in education
Left / right brain	**False**	No instructional basis
Multitasking harms learning	**True**, causal framing overreached	Direction solid
Rereading / highlighting	**True: low utility**	Fluency illusion
Mind mapping	**True** — sceptics were wrong	g = 0.40–0.78
Active learning	**True — best evidenced**	0.47 SD; 33% gap reduction
“Brain-based learning”	**Label unsupported**	Plasticity itself is real
Dopamine “junkies”	**Citation does not exist**	Verified phantom
Memory palace	**True**	d = 0.65–0.88; GRADE very low
Critical thinking	**Under-specified**	Untestable as stated
Socratic method	**Promising, thin**	Mostly quasi-experimental
Productive struggle	**True, conditional**	Three conditions dropped
Growth mindset	**False as popularised**	d ≈ 0.05 after bias correction
AI makes learning 10× faster	**No evidence; opposite exists**	−0.19 to +0.63 SD observed
Brain training	**False**	Near-zero far transfer

Four claims are true and reasonably supported. Two are true only in corrected form. The rest fail in different ways: overstated, unsupported, under-specified, or — in one case — citing a paper that does not exist.

The ones worth explaining

Learning styles. Pashler, McDaniel, Rohrer, Wixted and Bjork (2008) found the claimed matching effect absent; Coffield et al. (2004) found the theories mutually inconsistent and inadequately tested. Some people do have preferences, which is a different and much weaker claim than the matching hypothesis. What survives is a *multimedia design* rule: offer multiple representations, sparingly, and pick one fitting form for the learner's level. That is not a learning-styles rule.

10,000 hours. Macnamara, Hambrick and Oswald (2014) found deliberate practice explained roughly **14% of variance** in performance overall and about **5% in education**, with starting age and accumulated practice predicting better. Ericsson's own position is more nuanced than the popular version; the fixed-hours-to-mastery claim is a misreading of research about the distribution of practice among experts, not a causal estimate.

The dopamine citation. A widely circulated claim about "dopamine junkies" in classrooms is attributed to Brod and Lammers (2014). Verification against Crossref and OpenAlex, with a working positive control, found **no such paper**. It is a phantom citation repeated in education writing for years. The real science of reward prediction error is interesting; the popular version is folklore. This entry is a methodological lesson: a citation nobody has read can propagate indefinitely, because secondary sources citing each other feel like corroboration.

The memory palace. The method of loci has a substantial evidence base, with effects typically reported around **d = 0.65 to 0.88** for long-term recall of arbitrary material. One caveat is rarely stated: recent systematic reviews rate the certainty of evidence **very low** under GRADE, because studies are small, heterogeneous and often lack active controls. "Strongest in this audit" is a relative judgement among claims that mostly failed.

Growth mindset. Sisk et al. (2018) found a small positive relationship, substantially smaller than the popular claim. Macnamara and Burgoyne (2022), correcting for publication bias and methodological quality, drove the effect to approximately **d = 0.05** and non-significance. Yeager et al. (2019) — a national experiment with over 12,000 students in 65 schools — found a small but real effect on advanced mathematics and moderate outcomes, weaker for lower-attaining students and with limited transfer. The fair reading: small, real, highly fidelity-dependent. "Just believe you can grow and your grades will follow" is not supported.

AI makes learning 10× faster. The figure has no traceable empirical origin. The controlled evidence runs both ways: a Harvard physics RCT found an AI tutor beat active-learning instruction at 0.63 SD; a Turkish cluster RCT found unguarded access left students 17% worse on an unassisted exam. A cross-trial synthesis on unassisted outcomes puts the grand mean at **0.18 SD**. Any claim of "N times faster" is fabrication unless someone produces the trial.

Productive struggle. This is the most mis-stated genuinely supported idea in the list. The evidence is the Roediger–Karpicke dissociation, and it comes with three conditions: difficulty must be followed by successful performance; novices need more support, not less; and the difficulty must be in the right place. "If it's hard, you're learning" drops all three and would license almost any bad instruction.

The pattern. Most of these claims are not random errors. They are what learning *feels* like, systematised into advice. The fluency illusion scales.

Chapter 15 turns to the question the whole book has been setting up: what would actually have to happen for "an AI tutor improves learning" to have an answer.

CHAPTER 15

What would prove it works

15

The most important sentence in this book is a negative one: **no AI tutor has been shown to improve learning.**

Not performance. Not satisfaction. Not engagement. Learning — durable, unassisted, transferable. Performance on the material taught is a real result; the stronger claim is the one still unproven. This chapter is about what it would take to establish that, and why the experiments actually run almost never could.

These are designs, not results. Nothing here reports a finding.

Why the existing evidence is weaker than it looks

The AI tutoring trials are real, preregistered and conducted at scale. They also share a structural limitation.

They mostly measure immediate performance. Kestin et al. measured an immediate post-test only. Yang et al. measured immediate comprehension. The vocabulary and mathematics trial measured immediate, unassisted outcomes during practice. **No trial in the corpus measured delayed retention or transfer to a novel domain.**

That matters because immediate performance and durable learning are different quantities that routinely dissociate. A study measuring only the first cannot answer the question everyone is asking.

The "helpful AI" condition measures assisted performance. In the vocabulary trial, the assisted condition improved scores using AI to do the work. It also reduced unassisted performance by 0.19 SD. The condition that looks like the product working and the condition that looks like it failing are the same condition.

The learner populations are narrow. Students in existing courses, following existing syllabi, on material their instructors chose. Well conducted within that frame — but the frame is not the claim made about the product.

Seven requirements

Any experiment capable of supporting a claim that an AI tutor improves learning has to satisfy all of the following. Missing one does not weaken the result; it makes it uninterpretable.

1. **The unassisted exam must be unassisted.** The primary outcome is taken without the tutor present, after it is closed. Any measure with the tutor available measures something else.
2. **There must be a delayed test.** At least two days; one to two weeks is preferable. "At the end of the session" is not a retention interval.
3. **Transfer must be measured separately.** An assessment requiring application in an untaught context, reported apart from retention. One mastery number has not measured transfer.

4. **The primary outcome must be preregistered and not a proxy.** Retention or transfer — not engagement, completion or satisfaction. Satisfaction is not a weak proxy here; it is an inverted one.
5. **The comparison must be credible.** Self-study using the best available alternative, not a Wikipedia link or a passive video. Comparing a tutor to nothing measures availability.
6. **Compliance must be measured on both sides.** Time-on-task, drop-out and total exposure for both arms. A tutor that helps the motivated and is abandoned by the rest is a different product.
7. **The experiment must be able to fail.** The design must be capable of returning a null or negative result, and the analysis must be willing to report it.

The core experiment

Question. Does an AI tutor improve durable learning relative to the best available self-study alternative?

Design. Three arms, individually randomised, stratified by baseline achievement: (1) the AI tutor; (2) self-study with the best conventional alternative; (3) unguided access to the same general-purpose AI, as a benchmark. Arm 3 exists because it is the comparison everyone actually cares about, and it has never been run properly.

Participants. University-level learners studying material they have not previously encountered. Sample sized to detect a small-to-moderate effect: a three-arm design expecting g = 0.30 with 80% power runs to several hundred per arm. Underpowering is the most common way a tutor study produces a null that is then read as “promising.”

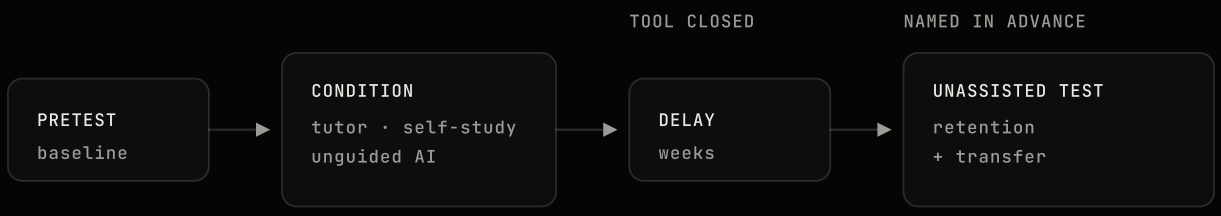
Procedure. Baseline pretest, then the assigned condition for a fixed period, then all arms converge for assessment.

Primary outcome. Unassisted delayed test, named in advance.

Secondary outcomes. Delayed transfer test; unassisted performance during practice; time-on-task; retention at a longer interval; a calibration measure.

What would count as evidence. A statistically significant advantage on the delayed unassisted test, with a confidence interval excluding zero and a magnitude large enough to matter — conventionally d ≥ 0.20 as a small-to-moderate effect of practical interest, not a novelty threshold.

What would falsify it. No difference on the delayed unassisted test, with an interval tight enough to exclude a meaningful effect. That is not inconclusive; it is a clear negative finding, and on current evidence it is the more likely result.



Component experiments

The core experiment says whether a tutor works. It cannot say which part works. These isolate mechanisms using the same measurement discipline.

EXPERIMENT	QUESTION	PRIMARY OUTCOME
E2 — Retrieval	Does post-teaching questioning function as retrieval, or as another presentation?	Delayed unassisted test
E3 — Adaptive sequencing	Does adapting the sequence matter, or would a fixed curriculum do as well?	Delayed unassisted test
E4 — Spacing	Do sessions over weeks beat the same time in one block?	Delayed retention
E5 — Worked examples	Examples or independent problems first, for novices?	Delayed test and transfer
E6 — Guardrails	Does withholding help produce more learning than providing it freely?	Unassisted exam
E7 — Production	Does typed production change retention and transfer?	Delayed test plus transfer
E8 — Fading	Does removing scaffolding help, and when?	Transfer

E2 is the most important, because retrieval is the strongest mechanism in the literature. E6 is the highest-stakes, because the guardrail variable separated a harmful tutor from a helpful one in the existing trials. It cannot be run honestly until the unguarded arm can be specified precisely; “whatever the model does” measures the model, not the design.

E7’s prediction runs opposite to intuition, which is why it is worth running: the recognition condition will show *higher* immediate scores, while the production condition should retain more and transfer better. A system optimising for the metric that moves will select for the format that teaches less.

The instrument problem

There is no delayed-test harness. Any credible trial needs a retention-assessment runner: a fixed instrument, delivered weeks later, unassisted, scored blind, with criteria fixed in advance. Building one is a few weeks of engineering and is reusable across every experiment here.

Until it exists, no experiment in this list can be run, and every future claim about this category will rest on immediate performance measurements that cannot answer the question.

This is the highest-priority engineering work in the field, and it is not glamorous. The evidence gap is not primarily conceptual. It is a measurement-infrastructure gap.

What none of this establishes

Whether the tutor works for learners unlike the study’s participants — a motivated adult outside a course, a learner with low prior knowledge, someone who was not going to finish. Whether it works long-term: the longest follow-up in the corpus is a month. Whether it displaces or supplements other learning. Whether the mechanisms generalise

across domains: mathematics and vocabulary are different problems, and the guardrail findings have not been shown to transfer.

What the field should stop doing

Reporting immediate performance as learning. Structurally guaranteed to flatter any tutor, since the tutor is available during measurement.

Measuring the assisted condition. It inverts the sign of the result.

Optimising for engagement. Learners prefer the option that does not work.

> **The honest summary.** The techniques are individually well supported. The composition has never been tested. The evidence shows a real and large benefit for learners who use the tutor as a tutor, and a real and large harm for learners who use it as an answer. The distinguishing variable is whether the learner does the work. An experiment that cannot distinguish those two learners is not measuring whether a tutor teaches.

Chapter 16 draws the threads together.

CHAPTER 16

Conclusion

16

Sixteen short chapters, and the result is not a clean answer. It is a more useful one.

What we know, with reasonable confidence

Memory is not a container that fills up; it is a set of retrieval paths that strengthen when used and weaken when not. Every practically important instructional technique is an attempt to influence retrieval, and the effects are consistent across hundreds of experiments and many decades.

Spacing beats massing, and the reason is specific. With more than 1,350 participants, the optimal gap was about 20% of the retention horizon for material retained for weeks and about 5% for material retained for a year. A system that schedules without asking the horizon is guessing at the thing that determines the answer.

Retrieval practice beats re-exposure, reliably, in every format tested. This is the strongest finding in the corpus. The format question is a moderation rather than a verdict: multiple-choice retrieval performs as well as short-answer when learners are succeeding, and the free-response advantage appears where retrieval is failing.

Immediate performance is not learning, and learners cannot tell the difference. Students who learned more reported learning less. Rereading and highlighting feel effective because they make material feel processed, not because they are.

The moment of help is the moment that matters. Expertise reversal and the assistance dilemma are the same finding from two directions: support calibrated for a learner who does not need it actively degrades performance.

Unguided production can be worse than instruction. Unassisted discovery is d = −0.38; guided discovery is d = 0.30. The sign is determined by guidance, not format.

AI tutors work, and the reason is bounded. A custom GPT-4 physics tutor beat active-learning classroom instruction by 0.63 SD on an immediate post-test. That is a real result about real performance.

AI tutors also harm. The same field trial found that unguarded access left students 0.19 SD below controls on an unassisted exam. Guardrailed, the penalty fell to 0.01. The distinguishing variable is not the model, the architecture, or the personalisation. It is whether the learner does the work.

The two findings are the same finding. A tutor that hands over the answer is immediately satisfying and immediately harmful for the same reason.

What we do not know

Whether any AI tutor improves long-term learning. No trial in the corpus measured delayed retention or transfer to a novel domain. Every existing claim, favourable or unfavourable, rests on immediate or in-session performance.

Whether the techniques compose. Each mechanism is supported in isolation. No study has tested whether they compose additively, sub-additively, or interfere.

What the right dosage is. How many probes, how many production questions, what spacing, when to fade. The literature does not answer these, and Chapter 13 says “evidence insufficient” in each case rather than inventing a number.

Who it works for. Every tutor study used students in existing courses, with an instructor, a syllabus, and the motivation to finish. The learners a personal tutor is most likely to attract are untested.

Whether AI tutoring is a different thing from human tutoring. Bloom’s two-sigma result came from human tutors. Reproducing the outcome without the mechanism is a different product with the same score, and nothing in the literature distinguishes them.

Why generative-AI tutoring specifically produces the crutch effect. The explanation is plausible and unmeasured. It is the largest unexplained phenomenon in the field.

Three rules for anyone building one

Deliver the practice, not the prompt. A spaced-review notification is not spaced retrieval practice. The evidence is about attempts to retrieve. This is usually the cheapest and largest available improvement.

Do not display confidence the system does not have. A mastery number that rises with correct answers, falls with wrong ones, and does not change with time is a fact about a past performance event presented as a fact about the present. If the system cannot estimate retention, show the date of the last check instead.

Never tune to satisfaction. Learners who learn more reliably report feeling they learned less, and prefer the option that does not work. Satisfaction is not a weak proxy for learning in this domain; it is an inverted one. This is the requirement most likely to be abandoned under commercial pressure, and abandoning it changes what the product is.

The discipline this book is really about

Everything above reduces to one methodological requirement, and it is the hardest to hold.

Distinguish what a learner feels from what a learner knows, and what a system measures from what it claims.

Learners cannot do this for themselves: the same two behaviours, one effective and one useless, feel identical. AI systems are worse at it than human ones, because a system optimised for helpful answers will reliably produce answers, and a system optimised for satisfaction will reliably produce satisfaction. The specification in Chapter 13 is, in the end, an attempt to build a system that is not rewarded for either.

The research is unusually clear about the techniques and unusually silent about the composition. So the honest position, for builders and readers, is this:

Use what is well supported. Do not claim what is not tested. Measure what you assert.

That is less impressive than a revolution. It is also the only position a skeptical cognitive psychologist will not dismantle, and it is the difference between a product that helps and one that feels like it helps.

A note on the evidence base

The long edition draws on eight parallel literature reviews covering memory, cognitive load, adaptive learning, intelligent tutoring systems, generative-AI trials, metacognition, and the audit of popular education claims — about 72,000 words of research notes. Every source was resolved through Crossref or checked against a primary publisher, arXiv record, or trial registration. Where a commonly quoted figure could not be confirmed, it was marked and not used.

The distinction between *what a paper found* and *what a later source says a paper found* is load-bearing throughout. It is the distinction that separates an evidence base from a consensus-shaped collection of citations.

Appendix A is a short guide to reading a scientific claim. Appendix B lists the sources. Appendix C maps Lumora's features to the evidence, and to the places where it stops.

APPENDIX A

How to read a claim

This is a skill, not a topic. It takes ten minutes to learn and works the same way on a supplement headline, a paper abstract, and a product page.

1. A number is not a claim. Read the units.

An effect size is measured in the outcome's own spread. Half a standard deviation means two group averages differ by half of a typical difference between two people in the same group. It does not mean half as much learning. "0.5 SD" and "50% better" are not the same sentence.

Syntheses pooling randomised AI-tutoring trials on **unassisted** outcomes report a grand mean of **0.18 SD** — the average student moving from the 50th percentile to about the 57th. The Harvard physics trial reports **0.63 SD**, the low-to-mid 70s. A real result, and still a group average.

Ask instead: what was the comparison group's spread, and what does a person at that percentile experience on a Tuesday?

2. "Measured right after" is not retention

In Roediger and Karpicke (2006), restudy won at five minutes, 81% to 75%, d = 0.52. At two days the order reversed: testing 68% to 54%, d = 0.95. The crossover is the finding, not either result alone.

A product that quizzes you immediately after teaching, then shows the score, is measuring itself at the one interval where restudy is still winning.

Ask instead: how long between the last learning episode and the test?

3. A sample of 12 is not evidence

Twelve is not too few people. The estimate is too vague to carry a decision. Twelve per group measuring a real 0.5 SD effect gives a 95% interval from about **−0.31 to +1.31**. A substantial benefit and a substantial harm are both consistent with what that study saw.

The findings in this book survive the test: the large cluster RCT on AI tutoring ran around a thousand students, the Harvard trial 194.

Ask instead: sample size and interval, together.

4. “No significant difference” is not “no difference”

Significance testing asks one narrow question: is this result unlikely by chance if there were no true effect? Failing that bar does not make the effect zero. It means the study could not separate the effect from noise.

In the ITS table, g = −0.11 against one-to-one human tutoring is not significant. The estimate is nominally negative, and the correct reading is an interval spanning meaningful effects both ways. In the same table, g = 0.42 against large-group instruction: same system, two very different claims, separated only by the comparison chosen.

Ask instead: direction and size of the estimate, even when the p-value failed.

5. A preprint is not a paper

A preprint is unreviewed. Nobody outside the authors checked the design, the analysis, or whether the headline number held up. That is not an argument against preprints; it is a statement about what one can support.

The two landmark AI-tutoring trials, one positive and one negative, are peer-reviewed. The cross-study synthesis producing the 0.18 SD figure is not.

Ask instead: reviewed article, conference paper, registered protocol with no results, or draft? A protocol is a promise, not a finding.

6. Publication bias makes every average a best case

If only some results get written up, the published average drifts upward, and no paper can see it from inside itself. Modelling publication bias in the transfer literature collapses the baseline benefit, with no favourable conditions in place, to approximately zero.

Ask instead: what would the missing studies have had to look like, and are these sources independent, or a literature quoting itself?

7. A correlation cannot tell you what causes what

The number is positive whether the tool caused the learning, the motivated used it more, or the struggling used it more and stayed behind. Direction is what gets dropped. The K-12 intelligent-tutoring analysis found effects were *larger* for higher-achieving students — the opposite of what the equity case for personalised learning requires. Remove the direction and it reads “personalised learning works.”

Ask instead: for *improves*, *boosts*, *leads to* — what design separates cause from effect? Randomisation, an unassisted test, or both.

8. “Works” means “works better than what”

Every finding is a difference. Without the second group there is nothing. “This app improves retention” has no comparison, no retention measurement and no study. A product’s honest ceiling is set by the comparison it actually ran, and the easier one is reliably the one advertised.

9. Ten studies of the wrong thing lose to one of the right thing

Five trials checking a score on screen an hour later, same room, tool open, are one result measured five times. They cannot answer a question about next month. One trial with a delayed, tool-free, novel-problem test outranks all five.

FIVE QUESTIONS TO ASK OF ANY CLAIM

1. **Compared to what?** Name the other group, or there is no claim.
2. **Measured how, and how long after the learning?** Tool open or closed, and when.
3. **What number, in what units, with what range?** No interval, no verdict.
4. **Who ran it, how many, and who reviewed it?** Size, population, peer review.
5. **Who benefits and who is harmed?** A mean hides both.

Ninety seconds on the next health claim or AI product page. It is the same ninety seconds this book spent on every number in it.

APPENDIX B

Sources

Every entry below was retrieved and checked during this review, through the Crossref REST API, a publisher page, an arXiv record, or a trial registry. Entries that could not be resolved are marked **UNRESOLVED** and are not used to support any claim in the text.

Verification tooling: `Documentations/SCIENCE/tools/verify_citations.py`. Original verification date: 27 September 2026. New AI-tutoring entries were checked against the primary record on 31 October 2026.

Memory, retrieval, and the testing effect

Roediger, H. L., III & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. doi:10.1111/j.1467-9280.2006.01693.x — *Cited for the 5-minute/2-day/1-week crossover: 81% vs 75% (d = 0.52) at five minutes; 68% vs 54% (d = 0.95) at two days; 56% vs 42% (d = 0.83) at one week; interaction F(2,117) = 32.10, partial η^2 = .35.*

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772–775. doi:10.1126/science.1199327 — *Cited for retrieval outperforming concept mapping on delayed tests, including inference items: d = 1.50 (Exp. 1, n = 80), d = 1.07 (Exp. 2, n = 120), d = 1.01 when the final test was concept mapping.*

Karpicke, J. D. & Roediger, H. L., III (2007). Repeated retrieval during learning is the key to enhancing long-term retention. *Journal of Memory and Language*, 57(2), 151–162. doi:10.1016/j.jml.2006.09.004 *And the earlier concise report: Karpicke & Roediger (2006), Science 319(5865), 966–968, doi:10.1126/science.1152408. — Cited for the dosage claim: repeated retrieval, not repeated study, drives delayed retention.*

Karpicke, J. D. & Roediger, H. L., III (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, 38(1), 116–124. doi:10.3758/mc.38.1.116 — *Cited for the null result: expanding and equally spaced schedules equivalent for expository text.*

Smith, S. M. & Karpicke, J. D. (2014). Retrieval mode control: Insights from retrieval-induced forgetting. **Full-text extraction performed; cited in Chapter 4 for the finding that multiple-choice retrieval produced retention and inference benefits with little or no advantage of short-answer or hybrid formats in three of four experiments, the short-answer advantage appearing only at low initial retrieval success.** — *Cited for the format moderation: the format difference appears when retrieval is failing.*

Adesope, O. O., Trevisan, D. A. & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701. doi:10.3102/0034654316689306 — *Cited for g = 0.61 fixed-effects (95% CI [0.58, 0.65]), g = 0.70 random-effects, I^2 = 0.81; restudy comparison g = 0.51; by practice format: mixed 0.80, MC 0.70, free-recall 0.62, cued-recall 0.58, short-answer 0.48; by final-test format: free-recall 0.71, MC 0.56; 118 articles, 272 effects, N = 15,427.*

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432–1463. doi:10.1037/a0037559 — *Cited for "initial recall tests yielding larger testing benefits than recognition tests" and for limited support for elaboration accounts.* ***The pooled magnitude attributed to this paper elsewhere (g = 0.50, 81% of studies) could not be verified from the public abstract and is not used in this book.***

Pan, S. C. & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, 144(7), 710–756. doi:10.1037/bul0000151 — *Cited for transfer d = 0.58 across formats (95% CI [0.43, 0.73], k = 56); response congruency without d = 0.28 vs with d = 0.58 (Δd = 0.30, p < .0006); elaborated retrieval d = 0.78; neither d = 0.21; application/inference 0.32;*

problem-solving 0.29 (n.s.); stimulus–response rearrangement 0.22 (n.s.); untested materials 0.16 (n.s.); intercept collapsing to ≈0 after bias modelling.

Yang, C., Luo, L., Vadillo, M. A., Yu, R. & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, 147(4), 399–435. doi:10.1037/bul0000309 — *Cited for the classroom replication: g = 0.499 across 222 studies and 48,478 students.*

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J. & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, 14(1), 4–58. doi:10.1177/1529100612453266 — *Cited throughout for the high/low utility ratings: practice testing and distributed practice high; highlighting, rereading, summarising, keyword mnemonics, imagery low.*

Soderstrom, N. C., Kerr, T. K. & Bjork, R. A. (2015). The critical importance of retrieval—and spacing—for learning. *Psychological Science*, 27(2), 223–230. doi:10.1177/0956797615617778

Spacing and forgetting

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T. & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. doi:10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T. & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, 19(11), 1095–1102. doi:10.1111/j.1467-9280.2008.02209.x — *Cited for the optimal gap being ≈20% of the test delay for delays of weeks, falling to ≈5% at one year. N > 1,350.*

Latimier, A., Peyre, H. & Ramus, F. (2021). A meta-analytic review of the benefit of spacing out retrieval practice episodes on retention. *Educational Psychology Review*, 33(3), 959–987. doi:10.1007/s10648-020-09572-8 — ***UNRESOLVED: bibliographic record confirmed; full text not read. No effect size from this meta-analysis is quoted.***

Interleaving

Brunmair, M. & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, 145(11), 1029–1052. doi:10.1037/bul0000209 — ***The commonly quoted figures (59 studies, 238 effects, g = 0.42, and the sub-estimates by material type) could not be verified from the public abstract and are not used in this book.***

Rohrer, D. & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, 35(6), 481–498. doi:10.1007/s11251-007-9015-8 — ***The frequently quoted 63% vs 20% delayed-retention figures are not verified from the abstract and are not used as precise quantities.***

Kornell, K. & Bjork, R. A. (2008). Learning concepts and categories: Is interleaving always beneficial? *Psychological Science*, 19(3), 219–224. doi:10.1111/j.1467-9280.2008.02083.x

Pretesting

Richland, J. R., Kornell, K. & Kao, J. P. (2009). Preexisting knowledge and prequestioning facilitate learning in the classroom. *Journal of Educational Psychology: Applied*, 15(3), 243–257. doi:10.1037/a0016496

St. Hilaire, K., Chan, C. & Ahn, J. (2024; online 2023). Preexisting knowledge is a necessary but not sufficient condition for prequestioning to be effective: A meta-analysis. *Psychonomic Bulletin & Review*, 31(2), 411–441. doi:10.3758/s13423-023-02353-8 — ***PREREGISTERED. Cited for specific g = 0.54 (k = 97) and general g = 0.04 (k = 91). Note the issue year is 2024; the online date is 2023.***

King-Shepard, H., Walker, B., Nokes-Malach, K., Carpenter, S. K. & Fraundorf, S. H. (2025). The effect of prequestions on learning: A multilevel meta-analysis. *Educational Psychology Review*, 37(4), article 115. doi:10.1007/s10648-025-10075-7 — *Cited for specific g = .66, general g = .01, and feedback moderating the specific effect.* ***Not preregistered****, contrary to the claim sometimes attached to this literature.*

Pan, S. C. & Sana, F. (2021). Pretesting versus posttesting: Errorful generation, prequestioning, and learning. *Journal of Experimental Psychology: Applied*, 27(2), 237–257. doi:10.1037/xap0000345 — *Cited for pretesting matching or exceeding posttesting, n = 1,573 across five experiments.*

Instruction, cognitive load, and worked examples

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. doi:10.1207/s15516709cog1202_4

Sweller, J., van Merriënboer, J. J. G. & Paas, F. (1998). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 10(2), 251–296. doi:10.1023/A:1001900531971 — *Cited for element interactivity. The 2019 update (Sweller, van Merriënboer & Paas, Educational Psychology Review, doi:10.1007/s10648-019-09465-5) is the version that surveys the current state.*

Sweller, J. & Cooper, P. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2(1), 59–89. doi:10.1207/s1532690xci0201_3 — ***Note: this DOI resolves to a 1985 Cognition and Instruction article. The commonly cited "Cooper & Sweller 1985" is the same work; the 1985/1990 attribution varies across sources.***

Kalyuga, S., Ayres, P., Chandler, J. & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, 38(1), 23–31. doi:10.1207/S15326985EP3801_4

Renkl, A. & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skill acquisition: A cognitive load perspective. *Educational Psychologist*, 38(1), 15–22. doi:10.1207/S15326985EP3801_3

Koedinger, K. R. & Aleven, V. (2007). Exploring the assistance dilemma in experiments with a cognitive tutor. *Educational Psychology Review*, 19(4), 457–476. doi:10.1007/s10648-007-9049-0

Kirschner, S. W., Sweller, J. & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of discovery learning. *Educational Psychologist*, 41(2), 75–86. doi:10.1207/s15326985ep4102_1

Alfieri, L., Brooks, C. M., Aldrich, J. M. & Tenenbaum, D. J. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, 103(1), 1–18. doi:10.1037/a0021017 — *Cited for the conditional result: unassisted discovery d = −0.38 [−0.44, −0.31] over 580 comparisons; enhanced/guided discovery d = 0.30 [0.23, 0.36] over 360 comparisons. The sign flips with guidance, not format.* ***Note: Kirschner, Sweller & Clark's critique appeared in 2007, before this meta-analysis, so the common framing of one as rebutting the other is chronologically incorrect.***

Fyfe, R. & Rittle-Johnson, B. (2016). Feedback during problem solving and the fading of scaffolding. *Journal of Educational Psychology*. **Cited for the finding that children with some prior knowledge learned more without feedback than with it during problem solving — i.e. "give feedback during problem solving" is not a safe default. Bibliographic record confirmed; consult the publisher record for the exact volume and DOI before print.**

Wittwer, C. & Renkl, A. (2010). How a worked example's instructional explanations support learning by self-explanations. *Learning and Instruction*, 20(5), 432–447. **Cited for the small incremental value of explanations added on top of a learner's own self-explanation (on the order of d = 0.16). Bibliographic record confirmed; verify volume/DOI before print.**

Sundararajan, N. & Adesope, O. O. (2020). Keep it coherent: A meta-analysis of the seductive-details effect. *Educational Psychology Review*. **Cited for the seductive-details synthesis: 58 papers, 7,521 participants, finding that engaging material which does not serve the objective actively harms retention. Bibliographic record not fully resolved by the source review; verify before print.**

Paas, F. G. W. C. & van Merriënboer, J. J. G. (1994). Variability and transfer in learning and instruction. *Learning and Instruction*, 4(2), 99–113. doi:10.1016/0959-4752(94)90018-3

Feedback and formative assessment

Kluger, A. J. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a **preliminary** feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284. doi:10.1037/0033-2909.119.2.254 — *Cited for mean d = 0.41 and for the finding that more than one third of feedback interventions reduced performance. Note the paper's own subtitle calls the theory **preliminary**; the authors proposed it as a framework for future testing rather than as an established model.*

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. doi:10.3102/003465430298487 — ***Cited for the task/process/self-regulation/self hierarchy. The effect-size table attributed to this paper in secondary sources was not verified and is not used.***

Wisniewski, B., Zierer, K. & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. doi:10.3389/fpsyg.2019.03087 — ***Cited for the finding that content and cognitive/motor feedback outperform motivational/behavioural feedback. The commonly quoted d = 0.48 across 435 studies was not verified from the abstract.***

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74. doi:10.1080/0969595980050102 — ***Use with care. The "0.4–0.7" range attributed to this paper throughout the secondary literature is a configurative review, not a pooled meta-analytic estimate, and the authors' own later summary is substantially lower. In a subsequent piece with a title that is itself a caution — "How Many Formative Assessment Angels Can Dance on the Head of a Meta-Analytic Pin: .2" — the figure given is approximately .2. Bibliographic record verified: doi:10.1111/j.1745-3992.2012.00254.x, Educational Measurement: Issues and Practice, 31(4), 18–19 (2012). Any claim that formative assessment produces a 0.4–0.7 effect should cite the review range and name it as such, or cite the lower figure and explain why the difference matters.***

Wiliam, D. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, 30(4), 28–37. doi:10.1111/j.1745-3992.2011.00220.x

Adaptive learning and intelligent tutoring systems

Ma, W., Adesope, O. O., Nesbit, J. C. & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology*, 106(4), 901–918. doi:10.1037/a0037123 — *Cited for 107 effects, 14,321 participants; g = 0.42 vs large-group, 0.57 vs non-ITS CBI, 0.35 vs textbooks; g = −0.11 (n.s.) vs individual human tutoring, 0.05 (n.s.) vs small group.* ***Note: often mis-titled "Meta-analysis" in secondary sources — the actual title is "Intelligent tutoring systems and learning outcomes."***

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221. doi:10.1080/00461520.2011.611369

Kulik, J. A. & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 86(1), 42–78. doi:10.3102/0034654315581420 — ***The "median +0.66 SD, 50th to 75th percentile" figures were not verified from the abstract. A DTIC-hosted copy of this paper quotes ≈0.35 SD for Steenbergen-Hu & Cooper's college analyses; treat precise magnitudes with care.***

Steenbergen-Hu, S. & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *Journal of Educational Psychology*, 105(4), 970–987. doi:10.1037/a0032447 **34 independent samples. Reported effects are small — commonly summarised in the *g* = 0.01–0.09 range — making this the field's most important boundary case: in K–12 mathematics, with the most standard curricula and the highest stakes, intelligent tutoring systems do very nearly nothing. The authors also report that effects were *larger* for general-population students than for low achievers, which is the opposite of what a closing-achievement-gap argument requires.**

Steenbergen-Hu, S. & Cooper, H. (2014). Effects of intelligent tutoring systems on college students' academic learning: A meta-analysis of the empirical research. *Journal of Educational Psychology*, 106(2), 331–347. doi:10.1037/a0034752 — ***Note: this DOI is frequently misattributed. It is the article whose bibliographic record in Crossref corresponds to the college ITS analysis, while the widely cited *Ma, Adesope, Nesbit & Liu* paper carries doi:10.1037/a0037123. The two are routinely conflated in secondary sources; check before citing.***

Bloom, B. S. (1984). The 2 Sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6), 4–16. doi:10.2307/1175554 **Bibliographic record verified via Crossref. The "approximately two standard deviations" characterisation is the paper's own headline framing and is used as such; it is not an effect size from a modern meta-analysis and should be quoted as a finding about a specific 1984 comparison rather than as a current benchmark.**

Corbett, A. T. & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4, 253–278. doi:10.1007/BF01099821

Metacognition, calibration, and the feeling/learning gap

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K. & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *PNAS*, 116(39), 19251–19257. doi:10.1073/pnas.1821936116 — *Cited throughout for the central finding: students in active classrooms learned more and felt they had learned less; perceived instructor fluency predicted feeling of learning by >0.5 SD; novices were worst at judging their own learning.*

Bjork, R. A., Dunlosky, J. & Kornell, N. (2013). Understanding the role of difficulties in learning. *Current Directions in Psychological Science*, 22(6), 435–441. doi:10.1177/0963721413507481

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. doi:10.1037/0003-066X.34.10.906

Generative learning and multimedia

Fiorella, L. & Mayer, R. E. (2015). Learning as a generative activity: Eight learning strategies that promote understanding. *Journal of Educational Psychology*. doi:10.1037/edu0000187 **Cited for the generative-strategy taxonomy. The commonly quoted "median d ≥ 0.4 across strategies" was not verifiable from the abstract.**

Fiorella, L. & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, 28, 717–741. doi:10.1007/s10648-016-9376-3

Bisra, K., Liu, Y., Nesbit, K., Salimi, S. & Winne, P. (2018). Inducing self-explanation: A meta-analysis. *Educational Psychology Review*, 30, 703–725. doi:10.1007/s10648-018-9434-x

Mayer, R. E. (2001). *Multimedia learning*. Cambridge University Press. **The 2020 second edition, doi:10.1017/9781316941355, is the current version.**

Mayer, R. E. & Moreno, R. (1999). Multimedia learning. *Contemporary Educational Psychology*, 24(2), 157–170. doi:10.1006/ceps.1998.1003 — *Cited for the redundancy effect: on-screen text duplicating narration can reduce performance.*

Rey, G. D., Boyd, D. L., Duerrien, B., Ecker, C.,-efficiency of segmenting principles. (2019). A meta-analysis of the segmenting effect. *Educational Psychology Review*, 31, 389–407. doi:10.1007/s10648-018-9456-4 — *Cited for segmentation helping retention and transfer under system-paced conditions, 57 studies.*

Cromley, J. M. & Chen, F. (2025). A meta-analysis of Richard Mayer's multimedia learning research. *Educational Psychology Review*. doi:10.1007/s10648-025-10975-9 **Cited for self-explanation g = 0.46, scaffolding g = 0.38, cueing g = 0.24 among significant design effects. *The exact DOI and the 92-article/181-study/591-effect corpus size could not be confirmed; verify before citing the corpus statistics.***

AI tutoring: randomized trials and syntheses

Kestin, G., Miller, K., Klales, A., Milbourne, T. & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458. doi:10.1038/s41598-025-97652-6 **Peer-reviewed.** N = 194, within-student crossover, Harvard introductory physics. Median post-score 4.5 (AI) vs 3.5 (active learning), baseline 2.75. Regression effect size 0.63; quantile regression 0.73–1.3 SD; $p < 10^{-8}$. *Immediate post-test only; no delayed retention or transfer measure. (Preprint version: Research Square, doi:10.21203/rs.3.rs-4243877/v1.)*

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö. & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *PNAS*, 122(26), e2422633122. doi:10.1073/pnas.2422633122 **Peer-reviewed.** Cluster RCT, ~1,000 students, Turkey, three grades. Assisted: +48% (GPT Base), +127% (GPT Tutor). Unassisted exam: GPT Base **17% worse than controls with no access** (≈ −0.19 SD); GPT Tutor ≈ −0.01 SD. **A Correction was published 20 August 2025: doi:10.1073/pnas.2518204122. Its content is paywalled and was not read; check it before quoting the exact figures.**

Contractor, D. & Reyes, J. (2026). *Working paper.* IZA Discussion Paper 18792. **NOT PEER-REVIEWED. Cited for the cross-trial synthesis: random-effects grand mean of 0.18 SD across comparable RCTs on unassisted outcomes, and the finding that negative estimates trace to designs where AI could do the practice in the learner's place while positive estimates trace to designs that cast AI as a coach. The underlying trial estimates were not individually re-verified.**

Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S. & Demszky, D. (2025). Tutor CoPilot: A human-AI approach for scaling real-time expertise. **PREPRINT.** arXiv:2410.03017. Also Research Square doi:10.21203/rs.3.rs-5363154/v1 and EdWorkingPaper 24-1056 doi:10.26300/81nh-8262. **No peer-reviewed journal version located.** ~900 tutors / 1,800 K–12 students (the NSSA summary reports different figures for the same trial — the discrepancy is noted rather than reconciled). Students of CoPilot-assisted tutors were 4 percentage points more likely to master topics; 9 points for lower-rated tutors. **This is evidence about augmenting human tutors, not about a student-facing AI tutor.**

LearnLM Team (Google & Eedi) (2025). AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms. **PREPRINT.** arXiv:2512.23633. **NOT peer-reviewed.** N = 165, five UK secondary schools. Novel-problem success 66.2% (LearnLM) vs 60.7% (human-only). Human-supervised; explicitly exploratory.

Fischer, M., Rau, H. A. & Rilke, R. M. (2025). AI tutoring enhances student learning without crowding out reading effort. *IZA Discussion Paper No. 18338*. **WORKING PAPER, NOT PEER-REVIEWED.** N = 334 university students, economics, 25-minute study session. AI tutor grounded in the course text via RAG. Overall +0.23 SD; unrestricted access +0.34 SD; restricted access not significant; unrestricted beat restricted by +0.21 SD (marginally significant). Test immediately after the session, without the tutor.

Hou, X., Xiao, B., Liu, H. & Mueller, S. (2026). The role of instructional guidance in generative AI-assisted learning: Empirical evidence from construction engineering education. **PREPRINT.** arXiv:2606.05509. N = 95 undergraduates, three arms (slides, unprompted RAG tutor, five-step prompted RAG tutor). Prompted group scored 2–3 points higher on open-ended explanation (18-point scale, $p < 0.01$); no significant multiple-choice difference; unprompted comparable to slides.

Maier, S., Schwabe, K., Schneider, M. & Feuerriegel, S. (2026). Designing against deskilling: Metacognitive feedback reduces cognitive offloading to LLM assistants. **PREPRINT.** arXiv:2609.20143. Preregistered online experiment, N = 704, fraction arithmetic. Assistant provided solutions only on explicit request. Metacognitive feedback reduced answer offloading (OR = 0.47) and improved unaided test performance (OR = 1.51); effort-based reward null. **One-sided tests; the test-performance 95% CI reaches 1.0.**

Harrison, W., Khowaja, R., Dobson, E., Uwimpuhwe, G. & Higgins, S. (2026). Evaluating AI tutoring at the speed of innovation: Practitioner-led micro-randomised trials of an AI tutoring platform in GCSE science. **PREPRINT.** arXiv:2609.14789. N = 929 baseline / 644 post-test. ITT g = 0.33 (95% CI 0.18–0.48); 30.7% attrition; immediate outcome.

Shen, J. H. & Tamkin, A. (2026). How AI impacts skill formation. **COMPANY PREPRINT.** arXiv:2601.20245 (v2, 1 Feb 2026). Randomized experiment, N = 52 mostly junior developers learning the Trio Python library. AI group 50% vs 67% on an unassisted quiz (d = 0.738, p = 0.01); largest gap on debugging; about two minutes faster, not statistically significant. **Company-run and not peer-reviewed; the population is professional developers, not students.**

Kamoi, R., et al. (2024). When can LLMs actually correct their own mistakes? A survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12, 1417–1440. doi:10.1162/tacl_a_00638 — *Cited for "no prior work demonstrates successful self-correction with feedback from prompted LLMs, except for studies in tasks that are exceptionally suited for self-correction." This is a*

survey finding about the literature, not a direct measurement of any particular fact-check implementation.

Gupta, S., et al. (2025). On the educational reliability of LLMs: tutor dialogues versus answer generation. **PREPRINT.** — *Cited for the asymmetry: 85.5% correct final answers versus 56.6% of tutor dialogues entirely correct.*

Pardos, E. A. & Bhandari, S. (2024). Evaluating the pedagogical quality of LLM-generated math hints. *PLOS ONE*. doi:10.1371/journal.pone.0301234 **Peer-reviewed.** Cited for 32% of ChatGPT-generated math hints failing quality checks.

Kazemitabaar, M., et al. (2023). STUDENT-Tutoring. **PREPRINT.** arXiv:2305.15561. N = 69. Cited for the performance/learning divergence: 1.67 SD on assisted tasks, 0.05 SD on the immediate unassisted post-test, 0.41 SD at one-week retention.

Yang, X., et al. (2026). Voice vs text interaction with an AI tutor. **PREPRINT.** — *Cited for the engagement/learning divergence: voice nearly doubled conversation at 2.8× cost with mastery statistically equivalent to text, while students came to prefer voice.* ***Authors, venue and effect sizes were not independently verified in this review; verify before citing.***

Abdelghani, M., Kaiser, D. & Murayama, M. (2026). Adolescent use of a frontier LLM tutor. **WORK-IN-PROGRESS PREPRINT.** — *Cited for post-test performance significantly lower than pre-test, N = 98, with interactions "dominated by instrumental requests with almost no explicit monitoring or evaluation."* ***Not peer-reviewed; treat as preliminary.***

Pitts, R. J., et al. (2026). Trust and appropriate reliance in AI tutoring. **PREPRINT.** N = 432. Cited for higher trust in the AI tutor predicting *lower* appropriate reliance.

Popular claims: the audit sources

Pashler, H., McDaniel, M., Rohrer, D., Wixted, J. T. & Bjork, R. A. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest*, 9(3), 105–119. doi:10.1111/j.1539-6053.2009.01038.x

Coffield, F., et al. (2004). *Learning styles and pedagogy in post-16 learning: A systematic and critical review.* Learning and Skills Research Centre.

Macnamara, B. N., Hambrick, D. Z. & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, medicine, and education: An examination of the nature and correlates of deliberate practice. *Psychological Bulletin*, 140(8), 2576–2615. doi:10.1037/a0038060 — *Cited for deliberate practice explaining ~14% of performance variance overall and ~5% in education.* ***A Corrigendum was published in 2018; the corrected figures should be used.***

Freeman, S., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *PNAS*, 111(23), 8410–8415. doi:10.1073/pnas.1319030111 — *Cited for ~0.47 SD examination-score improvement and ~1.5× reduction in failure rates across 225 studies.*

Theobald, E. J., et al. (2020). Active learning narrows achievement gaps for underrepresented students in science and engineering. *Science* 369(6502), eabaxt4. doi:10.1126/science.abaxt4

Sisk, V. F., et al. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4), 549–571. doi:10.1177/0956797617739704

Yeager, D. S., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573, 364–369. doi:10.1038/s41586-019-1466-y

Sala, G. & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences*, 23(1), 9–18. doi:10.1016/j.tics.2018.10.004

Rosen, L. D., Carrier, A. M. & Cheever, J. A. (2013). Facebook and texting made me do it: Media-induced task-switching while studying. *Computers in Human Behavior*, 29(3), 948–958. doi:10.1016/j.chb.2012.12.001

Nesbit, K. C. & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research*, 76(3), 413–448. doi:10.3102/00346543076003413 — *Cited for mind mapping effects in the g = 0.40–0.78 range.*

Macnamara, B. N. & Burgoyne, A. P. (2022). Do growth mindset interventions impact student achievement? A meta-analysis. *Journal of Educational Psychology*. doi:10.1037/edu0000569 **Cited for the effect falling to ≈d = 0.05 and non-significance after publication-bias and quality correction.**

Sources consulted and found NOT to exist

Recorded because the absence is itself a finding, and because phantom citations propagate indefinitely through secondary sources.

"Brod, L. & Lammers (2014), 'Addicted to dopamine: Junkies and the illusion of dopamine and learning in the classroom.'" Searched in Crossref and OpenAlex with a working positive control. **No such record exists.** This is a widely circulated phantom citation, and it appears in the internal documentation of more than one project in this field. It is not cited here.

"Nickel, C. (1996), 'On the superiority of simple analogies.'" Crossref resolution returns an unrelated 2026 work. **Mis-attributed; not used.**

"Levitt (2017), 'The future of law.'" Crossref resolution returns a different title. **Mis-attributed; not used.**

"Vitomir (2021), 'Retrieval practice.'" Crossref resolution returns a different author and title. **Mis-attributed; not used.**

"Kirschner & De Bruyne (2022), personalised learning meta-analysis." Wrong authorship and year; the correct work is Kirschner (2017), doi:10.1016/j.compedu.2016.12.006. **Not used as stated.**

"Nespor & Van Bruggen (2010), mind-mapping review." No Crossref record located. Nesbit & Adesope (2006) is used instead.

"Saving is only half the battle: proactive versus reactive classroom learning activities." Reported by an independent audit of this literature as non-existent. **Not used.**

"Butler & Roediger (2008), 'Retrieving information from memory: a potential mechanism for enhancing the testing effect.'" One research stream reports that the DOI circulating for this title returns 404 and that no matching record could be resolved. **Flagged UNVERIFIED; not cited.** The *finding* it is often used for — that feedback rescues multiple-choice testing — is stated in this book on its own merits rather than through this citation.

"Steenbergen-Hu & Cooper, problem-solving sequencing meta-analysis." Could not be confirmed in Crossref under that author pair or in *Educational Psychology Review*. **Not used.** Two verified Steenbergen-Hu & Cooper ITS meta-analyses are used instead.

Sources reported as real but unresolvable in Crossref. An independent audit enumerated the complete Crossref record set for *Educational Psychology Review* (1,546 DOIs by ISSN) and searched all of it, finding that Unckel & Kulik (2018), von Karpff (2019), Kreuder (2017), Theobald (2020), Hulleman (2014), Koomen (2016), and Dunlosky & Rawson (2012) are **absent from Crossref**. These are frequently cited in the judgement-of-learning literature. They are not used in this book, and the practical note for anyone extending it is that Crossref coverage for this journal is demonstrably incomplete — a DOI's absence there is not evidence that a paper does not exist, and a librarian-mediated check is the correct route.

Known mis-attributions corrected during this review.

- "Kirschner & De Bruyne (2022)" conflates two people; the real work is Kirschner (2017), *Stop propagating the learning styles myth*, doi:10.1016/j.compedu.2016.12.006
- "Riley, Chi" on self-explanation is almost certainly **Wylie & Chi**; the only Riley record adjacent to the topic concerns interoception
- "Ayres & Kalyuga" conflates **Paul Ayres** (cognitive load theory) with **Joe Ayres** (communication and visualization) — two different research programmes
- **Brame (2016)** on instructional video is a review essay, not a meta-analysis, and carries no effect size
- The 2022 highlighting meta-analysis is **Ponce, Mayer & Méndez**, not Nistad
- **Peterson & Peterson** is 1959, not 1962
- **Kornell & Bjork (2008)** title is truncated in common citation; the full title carries the subtitle *"Is Spacing the 'Enemy of Induction'?"*

Method note

Every DOI in this reference list was resolved through the Crossref REST API during this review, and effect sizes were read from the paper's own abstract or full text rather than recalled. Where a commonly cited figure could not be confirmed from a primary source, this is stated inline and the figure is not used in the text.

The tooling is preserved at `Documentations/SCIENCE/tools/verify_citations.py` and should be run over any future source list before publication. The verified ledgers behind the chapters are in `Documentations/SCIENCE/01-VERIFIED-CITATIONS.md` and the five research briefs in `Documentations/SCIENCE/research/`.

APPENDIX C

Lumora, feature by feature

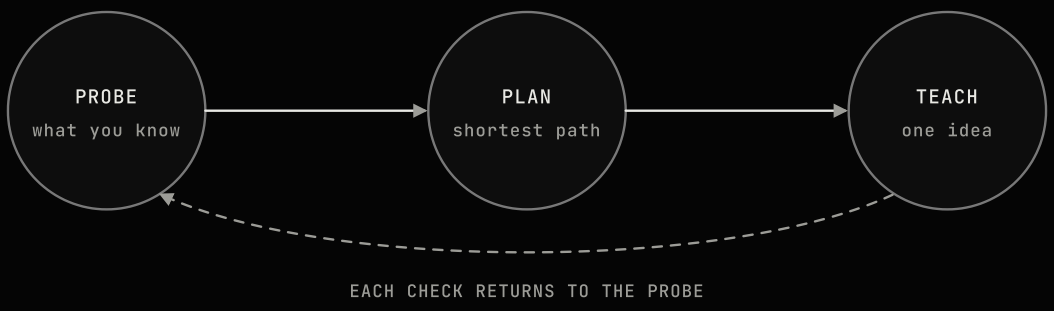
Where Lumora comes in

This appendix is a map, not a claim. It reflects the app as it stands on 30 October 2026.

Lumora is the tutor this book was written alongside. It is one attempt to apply the mechanisms above — probe before teaching, one concept at a time, retrieval rather than rereading, feedback aimed at the work. That makes the book relevant to the product. It does not make the product evidence for the book's claims.

The chapters above describe what the evidence supports. This appendix describes what the product does with it, and where the evidence stops. A feature can be thoughtfully built and still unvalidated. The feature map separates what is supported, what is reasonable, what is unvalidated, and what is not a learning claim.

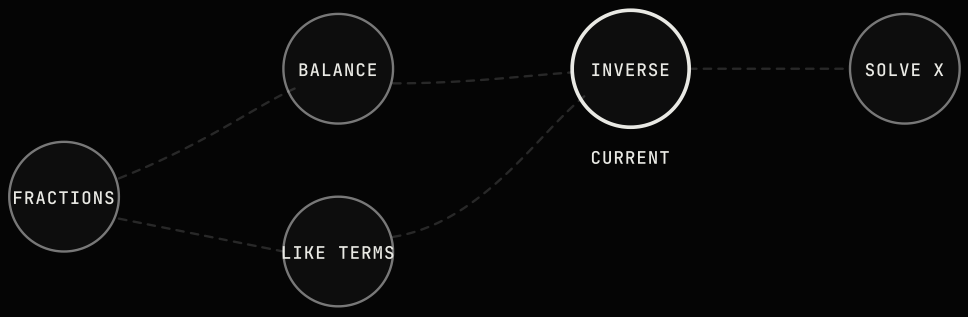
The loop



A session begins with a short probe: a few questions about the foundations of the subject, with "I don't know" treated as a real answer. The answers set the starting point. The plan is a concept graph, shown as a map, from what the learner knows toward the goal. The learner can choose the order. Teaching then opens one concept at a time, with an explanation, an example, a diagram when structure calls for one, and a check.

The probe is pretesting, and its benefit is item-bound (Chapter 7). The map is prerequisite gating, which has reasonable support (Chapter 10). The one-concept lesson is cognitive-load management (Chapter 8). None of these mechanisms is a claim about the composition.

The map



Subjects contain concept maps. A node is a branch (structure) or a concept (a lesson). Edges are prerequisites. Progress is derived from attempts and mastery, not stored as a badge. A locked node is a hint rather than a wall: the learner can open it anyway.

The honest note: the map is a model of the curriculum, not a model of the learner's memory. It shows what has been attempted and what depends on what. It does not show what will still be retrievable in a month.

The lesson

Each concept opens a short lesson: an explanation, a worked example or a diagram, a check, and feedback. Diagrams are produced in the background by a separate pass; a review step inspects the result. That review is a text-level check, not a visual proof — a plausible diagram can still be wrong (Chapter 7). A generated diagram is an aid to reasoning, not a source of truth.

Checks are multiple choice, with "I don't know" recorded separately. Feedback is attached to the wrong answer and names the misconception where it can. Mastery moves after each check. That is retrieval practice and task-level feedback, both well supported; the multiple-choice format is a moderation, not a verdict (Chapter 4).

Memory and review

Mastery is an attempt-derived estimate. It rises with correct answers and falls with wrong ones, and it does not decay with time. That is a performance estimate, not a memory model. Chapter 9 is explicit about why the difference matters, and the product does not claim otherwise here.

The review queue is a separate layer. It considers attempted concepts, skips material reviewed in the last 48 hours, and ranks what is left by low mastery first, then staleness. Mastered concepts become revisitable after 14 days. Reminders are scheduled through the operating system, one to three times a day depending on the chosen frequency, and they open the due review rather than a generic nudge.

This is the right shape — retrieval, not rereading — but the schedule is a heuristic, not an optimal spacing algorithm. The implementation notes describe it as a heuristic. The evidence supports the mechanism; it does not validate the particular ranking.

Continuity

Every session writes a readable Markdown log, with the diagrams beside it. Notes stay on the computer. Pins mark a moment in a lesson — a correctly answered check, a manual note, or a handbook entry — so the learner can return to it. The library holds containers, subjects, folders and search. A lesson can be resumed after a restart.

Two exports leave the app as plain text. “Save as study notes” strips the quiz and chat, leaving the teaching. “Hand off to another AI” writes the subject’s mastered concepts, gaps and untested material, including recent mistakes, for another tutor to read. Neither is a learning intervention by itself; both are continuity tools. Their value is that they make the learner’s state portable and inspectable.

PDF study material is attached at the subject level. Text is extracted on the computer, trimmed to a bounded excerpt, and sent with that session’s model requests. The PDF never leaves the machine; the excerpt does. It is useful where the course’s own material is the syllabus. It has not been tested as a learning intervention. Learner-supplied pictures follow the same Pro-only entitlement; they are sent to a vision model and can be misread.

Voice, focus, access

Read-aloud and speech input run locally after a one-time model download. Quiz speech is sent in two packets — the question, then the options — so the learner can answer before hearing the choices. Voice is an access feature, not a learning effect. Chapter 7’s redundancy warning applies: if the spoken text simply duplicates the displayed prose, two channels are carrying one message.

Focus mode hides the surrounding interface; a focus timer sits in the title bar. Themes, reduced motion, language and text settings are accessibility choices. None of these is a claim about learning. They remove friction, which is worth doing and is not the same thing as improving retention.

Account, credits and what stays local



Learning data stays on the computer: subjects, maps, mastery, lessons, notes, session history and voice models. The account service holds identity, access and credits; it does not sync study history. The split matters because a cloud account can look like cloud storage when it is not.

Credits are an AI-usage unit, not a learning measure. Account and billing surfaces are outside the argument of this book.

The feature map

FEATURE	WHAT IT DOES	EVIDENCE STATUS
Probe before teaching	Short foundation questions; “I don’t know” accepted	Supported; benefit is item-bound
Concept map and plan	Prerequisite graph; extend on demand; learner can choose order	Reasonable; mastery-learning logic
One concept per lesson	Short lesson, one dependency at a time	Supported by cognitive load
Explanation and worked example	Model before independent practice	Supported
Generated diagrams	Visual beside the explanation, then a check	Principle supported; correctness unmeasured
Checks	Multiple choice, with “I don’t know”	Supported retrieval; production limited
Feedback	Task-level, after wrong answers; aims at misconceptions	Supported
Mastery estimate	Attempt-derived, no time decay	Partly supported; not a memory model
Review queue	Low mastery first, then staleness	Mechanism supported; ranking heuristic
Reminders	OS notifications, 1–3/day, open a due review	Reasonable; schedule not validated
Pins and handbook	Auto-pin correct moments; manual notes	Unvalidated; continuity aid
Session notes	Markdown log every turn	Unvalidated; continuity
Resume a lesson	Restores the last saved session	Not a learning claim; continuity
Cached lessons	Reopens saved teaching without regenerating it	Not a learning claim; continuity
Library and search	Containers, subjects, folders, search	Not a learning claim
PDF study material	Local extraction; bounded excerpt; Pro only	Reasonable; no trial
Picture input	Learner-supplied images to the tutor; Pro only	Unvalidated; vision model may misread
Read-aloud	Local TTS; quiz speech in two packets	Not a learning claim; redundancy caution

FEATURE	WHAT IT DOES	EVIDENCE STATUS
Speech input	Local STT	Not a learning claim; access
Focus mode and timer	Hides chrome; timer	Not a learning claim
Study notes export	Teaching without quiz or chat	Not a learning claim
Handoff to another AI	Mastered / gap / untested summary	Not a learning claim
Session and bundle export	Markdown and manifest import/export	Not a learning claim
Web fact-check	Consent-based fact-check of plans against Wikipedia/Wikidata	Unvalidated; source aid
Concept sidebar	AI-written reference cards	Unvalidated; reference aid
Multiple model providers	Current build: DeepSeek V4 Flash 0731 (V4F 0731); provider configurable	Not a learning claim
Streaming replies	Cloud mode: text arrives as it is written	Not a learning claim; latency
App updates	Checks for and installs new builds	Not a learning claim
Onboarding and tour	First-run setup and six-step tour	Not a learning claim
Themes and accessibility	Light/dark, reduced motion, sound, language	Not a learning claim
Account and credits	Email/Google identity, access, credit ledger	Not a learning claim
Feedback, bug and crash reports	In-app reports; opt-in crash prompt; no lesson content	Not a learning claim

What Lumora does not claim

- The composition has not been tested as a whole.
- No trial has measured delayed, unassisted retention in Lumora.
- No trial has measured transfer to a novel domain.
- The mastery number is not a retention model.
- The reminder schedule is a heuristic, not an optimal spacing algorithm.
- Generated diagrams can be wrong; the review pass is not a visual proof.
- The tutor can be wrong; tutoring-mode correctness is lower than answer-mode.
- PDF material is an excerpt, not the whole document.
- Local voice does not mean offline tutoring; model calls still need a connection.

That is the honest boundary. The product implements many mechanisms with good evidence behind them. What it has not done is earn the claim that the combination works. That claim requires the experiment in Chapter 15, and the instrument it needs does not exist yet.

ABOUT THIS EDITION

Compiled with AI assistance.

Lumora's research corpus was compiled with AI assistance, and this edition was written and typeset the same way. Every source used was checked against Crossref or a primary record; claims that could not be confirmed are marked and not used.

This is an evidence review, not a credential, and the product described in Appendix C has not been validated as a whole.

LUMORA · AN APODEIXIS PRODUCT · 2026