



Cognitive Architecture and Digital Learning in Higher Education

A Research-Backed Analysis of Course-Integrated Study Systems

David Wild¹ Kartikey Dharendra²

¹Gradarina ²University of Liverpool

Version 1.0 • 19th September 2026



Abstract

Higher education faces two simultaneous pressures. The first is the *help-seeking paradox*: the students in greatest academic jeopardy are systematically the least likely to approach the support services designed for them. The second is a technological shift in which general-purpose generative AI, while lowering the social cost of asking a question, encourages passive study habits that inflate confidence without improving comprehension.

This paper examines the cognitive, behavioural and structural foundations of a course-integrated study system that ingests primary institutional artefacts — virtual learning environment exports, lecture slides and video transcripts — and converts them into active study material. We evaluate four mechanisms against the empirical literature: interaction-bounded segmenting, automated retrieval gates, a three-tier question-generation pipeline, and aggregated cohort-level feedback.

We find strong empirical support for retrieval gating, friction removal in support delivery, and aggregate normative feedback. We find the literature actively *contradicts* two design decisions: approximately thirty-second lecture slicing, and the treatment of prompt-constrained question tiers as calibrated difficulty. We state both plainly rather than defending them, on the view that a position paper which cannot name its own weak points is not evidence but marketing.

Contents

1	The Help-Seeking Dilemma and Institutional Support Delivery	3
1.1	The Help-Seeking Paradox	3
1.2	Opt-In Architectures versus Automated Course Ingestion	4
2	Generative AI, the Fluency Illusion and Metacognitive Degradation	4
2.1	Illusions of Competence and Metacognitive Laziness	4
2.2	Instrumental versus Executive Help-Seeking	5
3	Retrieval Practice, Desirable Difficulties and Testing Mechanisms	6
3.1	The Testing Effect and Memory Consolidation	6
3.2	The Hypercorrection Effect and Progress Framing	6
3.3	Known Shortcomings and Implementation Limits	7
4	Segmenting and Cognitive Load: The Fragmentation Debate	7
4.1	Cognitive Load Theory and the Segmenting Principle	7
4.2	Where the Literature Contradicts the Design	8
4.3	Preserving Coherence Through Knowledge Graph Topology	8
5	Question Difficulty, Bloom's Taxonomy and Assessment Rigour	9
5.1	The Low-Bloom Clustering Problem	9
5.2	Interpretable Difficulty via Knowledge Graphs	9
5.3	Prompting versus Calibration: The Honest Distinction	10
6	Social Dynamics, Motivational Architecture and Psychological Safety	10
6.1	The Pitfalls of Public Cohort Feeds	10
6.2	De-Stigmatising Struggle Through Aggregate Insight	10
6.3	Autonomy in Gamification Architecture	11
7	Synthesis: The Platform Against the Evidence	12
7.1	What Follows	13
A	Citation Audit	14
A.1	Corrections applied to the source draft	14
A.2	Still requiring verification	14
A.3	Sections where the evidence is thinnest	15

1 The Help-Seeking Dilemma and Institutional Support Delivery

The rationale for automating study support directly from course artefacts rests on the psychological mechanics of academic help-seeking. In higher education, help-seeking is not a rational calculation of cognitive need. It is a social transaction carrying significant psychological cost [1, 2].

1.1 The Help-Seeking Paradox

Empirical research across post-secondary institutions demonstrates that academic help-seeking runs inverse to academic need: the learners requiring the most scaffolding are systematically the least inclined to seek it [1, 3]. Karabenick and Newman's self-regulated learning framework explains this through ego threat and social-evaluative anxiety [2]. Admitting academic difficulty requires public vulnerability, which students interpret as conceding low inherent ability rather than taking an adaptive step toward mastery [1].

This avoidance concentrates sharply within particular cohorts:

- **Low academic self-efficacy learners.** To protect self-worth, struggling students conceal difficulty from peers and faculty [1, 3].
- **First-generation and underrepresented students.** Heightened stereotype threat and impostor phenomenon lead support requests to be read as confirmation of institutional non-belonging [2, 4].
- **Performance-avoidance oriented individuals.** Students whose achievement goal is avoiding the appearance of incompetence show the highest help-avoidance rates [2, 5].
- **Students socialised to traditional masculine norms.** Lower help-seeking among male students is mediated by internalised self-reliance scripts that frame collaborative support as a loss of autonomy [6, 7].

Struggling students additionally suffer metacognitive monitoring failures, remaining unaware of conceptual deficits until summative failure [8].

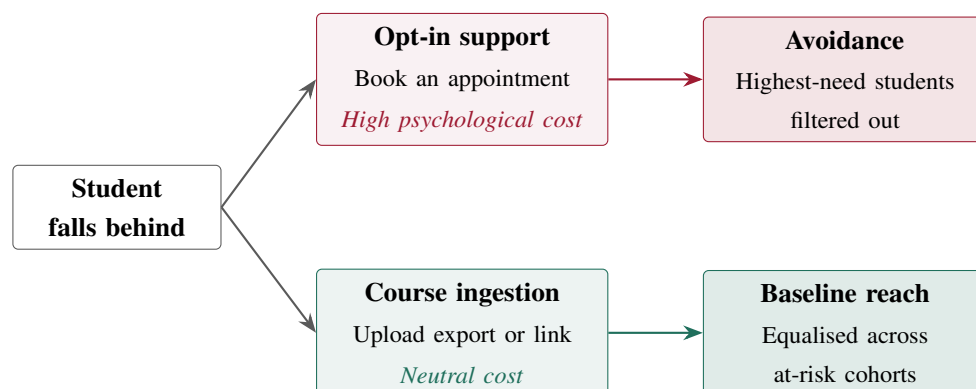


Figure 1: Conceptual schematic (not empirical data). The psychological cost of the support trigger determines which students the pathway reaches. Opt-in architectures filter *against* the cohort they are designed to serve.

1.2 Opt-In Architectures versus Automated Course Ingestion

Traditional higher education support is built around voluntary, student-initiated models: office hours, drop-in tutoring, counselling portals. As Tinto established, opt-in designs fail to reach non-retained cohorts because they require students to overcome administrative friction, scheduling anxiety and perceived stigma before any support is delivered [4]. Traditional services therefore disproportionately serve already-confident, mastery-oriented students, widening rather than narrowing the attainment gap [2, 4].

Table 1: Support delivery architectures compared across four dimensions.

Dimension	Student-initiated (opt-in)	Course-integrated ingestion
Operational trigger	Student admits struggle, books an appointment, visits faculty	Student uploads a course export or lecture link
Psychological cost	High — requires public vulnerability and exposure of deficiency	Neutral — automated, private, decoupled from human judgement
Administrative friction	Multi-step navigation across disparate services and schedules	Single-action ingestion generating lecture-specific material
Population reach	Skews toward high-efficacy, socially resourced students	Equalises baseline access across enrolled cohorts

The system's onboarding — importing a course export, lecture transcript or video link — operationalises Earl's intrusive advising philosophy [4, 9]. By converting a student's actual course materials into pre-scaffolded study suites with near-zero setup friction, it sidesteps the staged administrative and psychological hurdles that inhibit help-seeking [1, 9].

Constraint

This argument establishes that ingestion *removes a known barrier*. It does not establish that removing the barrier improves attainment. That claim requires outcome data from a real cohort, which this paper does not yet have. The mechanism is evidenced; the effect is hypothesised.

2 Generative AI, the Fluency Illusion and Metacognitive Degradation

The integration of commercial large language models into university study habits has lowered the social-evaluative barrier to asking questions. Interacting with a machine carries no fear of interpersonal judgement, allowing introverted and at-risk students to query foundational concepts freely [8, 10]. Post-2022 research reveals an equally severe cognitive hazard in exchange: the **fluency illusion**.

2.1 Illusions of Competence and Metacognitive Laziness

Cognitive psychology has long shown that the subjective ease with which information is processed inversely correlates with long-term retention and transfer [12, 13]. As Koriatic and Bjork established, when a worked solution is immediately visible, learners experience an illusion of competence: they mistake the ease of *following* a polished solution for the ability to *reproduce* it independently [12].

Fan et al. characterise the downstream behaviour as *metacognitive laziness*: with unconstrained access to generative AI, students reduce internal self-monitoring, accept fluent algorithmic answers at face value, and bypass the reflective evaluation needed to detect gaps in their own mental models [11]. Yan et al.'s scoping review of large language models in education identifies the same pattern among the field's recurring practical and ethical risks [14].

2.2 Instrumental versus Executive Help-Seeking

Self-regulated learning theory distinguishes two forms of help:

- **Instrumental help-seeking.** Seeking the minimum scaffolding, hint or clarification required to reach independent mastery [1, 2].
- **Executive help-seeking.** Asking another agent — human or machine — to execute the task directly, bypassing the cognitive work [1, 14].

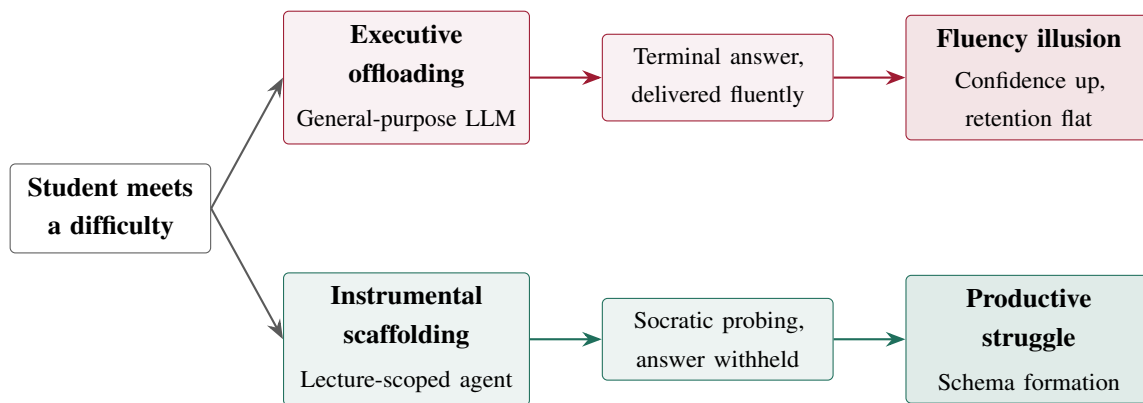


Figure 2: Conceptual schematic (not empirical data). The same student difficulty resolves into opposite cognitive outcomes depending on whether the agent supplies the answer or withholds it.

General-purpose chatbots are trained to optimise for speed, user satisfaction and conversational agreeableness [14]. Presented with a complex academic problem, their default is a complete, articulate terminal answer. This converts the learning transaction into executive offloading, encouraging students to skip precisely the stages where durable schema formation occurs [8, 14].

The system addresses this through a lecture-scoped conversational agent bound to the uploaded materials and constrained to Socratic guidance rather than terminal answers. The agent also runs a distress protocol that suspends tutoring and surfaces verified crisis resources when emotional distress is detected.

Constraint

The Socratic constraint is a prompt-level and retrieval-level restriction, not a guarantee. Any sufficiently determined student can obtain a terminal answer by rephrasing, and the constraint has not been adversarially evaluated. Claiming otherwise in front of an academic audience would be indefensible.

3 Retrieval Practice, Desirable Difficulties and Testing Mechanisms

To prevent passive consumption and break the fluency illusion, the system embeds automated testing: flashcards with Leitner-style spacing, lecture-level quizzes, and multi-chapter quick-checks.

3.1 The Testing Effect and Memory Consolidation

Decades of empirical work identify active retrieval as among the most effective study strategies available [15, 16]. Roediger and Karpicke demonstrated that pulling information *from* memory alters and consolidates memory representations far more effectively than repeated passive review [15].

Dunlosky et al. evaluated ten widely used learning techniques and assigned each a utility rating based on generalisability across materials, learners, and criterion tasks. The result, shown in Figure 3, is uncomfortable for conventional study habits: the two techniques students most commonly use — highlighting and rereading — both landed in the lowest band [16].

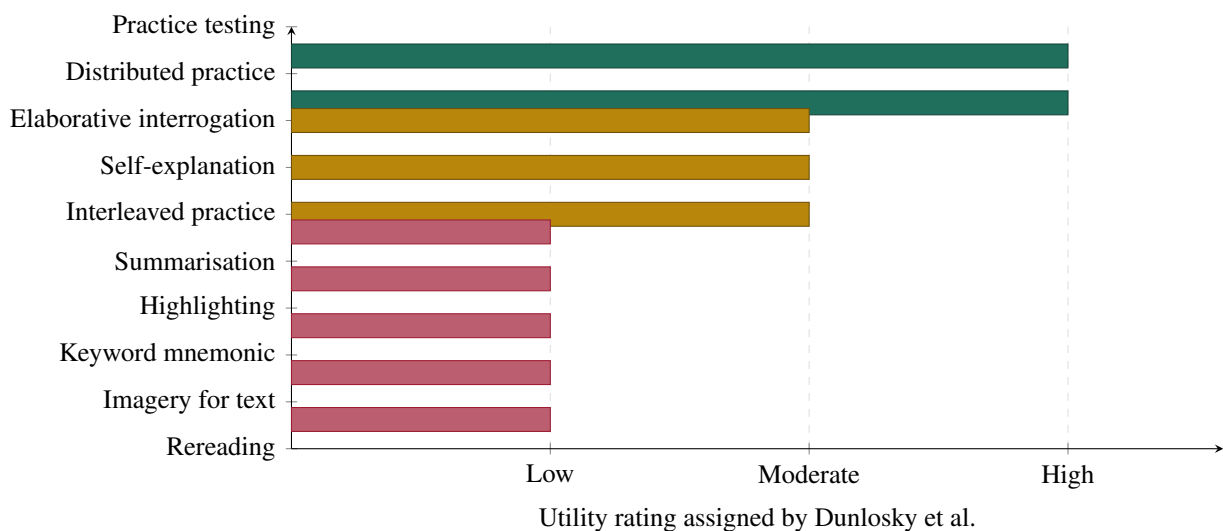


Figure 3: Utility ratings for ten learning techniques [16]. Ratings are **ordinal categories assigned by the authors**, not effect sizes — bar length encodes rank only and should not be read as magnitude. Practice testing and distributed practice are the two techniques the system’s quiz gates and flashcard scheduling target directly.

The system operationalises this by placing mandatory **quiz gates** between video chapters and before broader summaries unlock. This enforces what Bjork terms a *desirable difficulty* [13]: requiring effortful retrieval before new material disrupts the passive flow of video consumption and recalibrates the student’s metacognitive self-assessment [12, 15].

3.2 The Hypercorrection Effect and Progress Framing

The quick-check mechanism tracks incorrect responses and re-serves them on later passes with explicit progress feedback of the form “*you got this wrong before — right this time*”.

This reflects the **hypercorrection effect** [17]. When a student makes an error held with high confidence and receives immediate corrective feedback, memory updating on re-test is significantly

greater than after a marginally correct initial response [17]. Signalling the correction explicitly reinforces student agency and frames struggle as the mechanism of mastery rather than evidence against it.

3.3 Known Shortcomings and Implementation Limits

Where the architecture falls short of the literature

The delayed spacing gap. The testing effect depends heavily on temporal distribution across days [16]. The system prioritises previously-wrong answers on the *next immediate attempt*, but has no cross-session scheduler forcing retrieval intervals over days or weeks. Without temporal distribution, repeated testing inside a single session yields diminishing returns [13, 16].

The binary feedback limitation. Corrective feedback must carry explanatory elaboration anchored to the underlying concept [16, 17]. Where quizzes give only correct/incorrect confirmation without rationale, the hypercorrection effect weakens and students memorise answer keys rather than principles [8, 17].

4 Segmenting and Cognitive Load: The Fragmentation Debate

A central feature of the platform's short-form feed is automated parsing of uploaded lectures into vertically scrolling chapters interspersed with quick-check items. This sits directly on a live fault line in the learning sciences: the balance between reducing working-memory overload and preserving conceptual coherence [18, 19].

4.1 Cognitive Load Theory and the Segmenting Principle

Working memory is severely constrained, able to process only a small number of novel interacting elements at once [18, 20]. Uninterrupted lecture instruction frequently induces cognitive overload when pacing outstrips the student's capacity to organise and integrate [18].

Mayer's segmenting principle shows that dividing continuous presentations into discrete, learner-paced units improves comprehension and transfer [19, 21]. The mechanism is pacing: pauses between units give working memory the bandwidth to consolidate before new elements arrive [19, 20].

4.2 Where the Literature Contradicts the Design

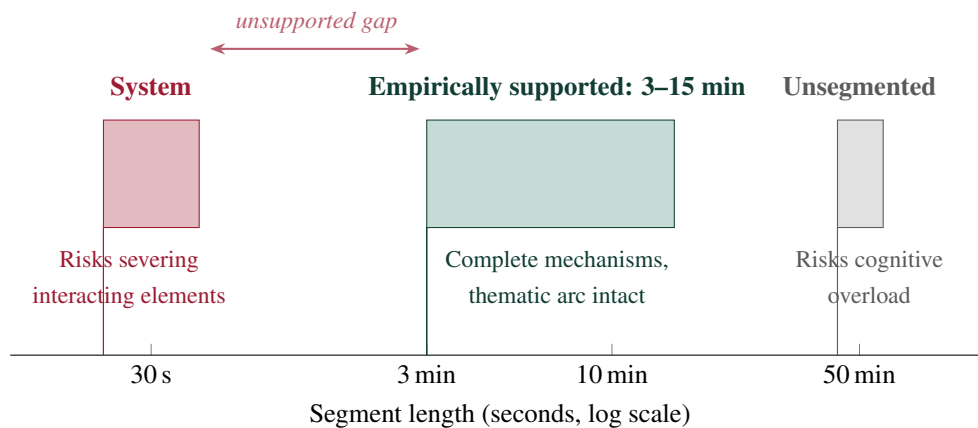


Figure 4: Segment length against the empirically supported band. The supported range is drawn from higher-education segmenting research [19, 21, 22]; the system’s position is its current implementation. **The gap is roughly an order of magnitude, and it is not supported by evidence.**

The system’s slicing of lecture recordings into segments of approximately thirty seconds is an extreme extrapolation of the segmenting principle, and the literature challenges it on three grounds.

1. **Absence of empirical precedent.** Higher education research supports segmenting into roughly ten-to-fifteen-minute conceptual blocks to sustain vigilance and permit re-entry after attention lapses [19, 22]. There is no robust evidence that atomising university-level discourse into thirty-second bursts improves structural understanding [21, 22].
2. **Severing conceptual coherence.** Complex subjects — multivariable calculus, organic synthesis, macroeconomic equilibrium — are characterised by high element interactivity [18, 20]. Chopping a continuous causal chain into thirty-second fragments discards the connective tissue between concepts [19, 23]. Kalyuga’s work on schema formation indicates that isolating interactive elements into sub-functional fragments impedes construction of high-level mental models, leaving students with isolated definitions rather than systemic comprehension [20, 23].
3. **Habituating superficial consumption.** Presenting lecture content as a vertically scrolling, mobile-first feed risks training rapid, superficial swiping. A friction-free interface conflicts directly with the principle of desirable difficulties, manufacturing an illusion of progress as students move through cards without processing the instructional narrative [13].

4.3 Preserving Coherence Through Knowledge Graph Topology

To reconcile segmenting with structural coherence, the literature emphasises explicit hierarchical roadmaps [23, 24]. The system incorporates an automated concept map extracting nodes and relational edges across the uploaded syllabus.

Structural concept maps act as an *advance organiser*, helping students locate a micro-segment within the domain hierarchy [23, 24]. When segments are anchored to a visible directed acyclic graph, students can bridge a specific instructional step back to its prerequisites and forward to its applications, partially counteracting the fragmentation that short-form delivery introduces [19, 24].

The honest position on segmenting

The concept map is a *mitigation*, not a refutation. It does not convert an unsupported thirty-second slice into a supported one; it provides a structural index alongside it. The defensible route is to treat segment length as an empirical question the pilot can answer, and to instrument the cohort so it does.

5 Question Difficulty, Bloom's Taxonomy and Assessment Rigour

The system generates three tiers of assessment item from course materials — **Recall**, **Apply** and **Analyse** — reflecting faculty demand for assessment targeting higher-order cognition.

5.1 The Low-Bloom Clustering Problem

Automated question generation driven by large language models exhibits a documented systemic flaw: items cluster overwhelmingly in the lowest tiers of the revised Bloom taxonomy, *Remember* and basic *Understand* [14, 25]. Prompted to generate “hard” or “analytical” questions using natural language alone, models frequently produce pseudo-complexity: lengthy vignettes that still assess vocabulary identification [14, 25]. Faculty scepticism toward automated quiz tools is largely grounded in this, since university curricula demand relational trade-offs, causal diagnosis and cross-concept synthesis [25, 26].

5.2 Interpretable Difficulty via Knowledge Graphs

Şakiroğlu, Güvenir and Kaya demonstrate that question difficulty can be modelled and interpreted through knowledge graph topology rather than black-box prompt engineering, computing difficulty signals from graph structure and combining them into a unified, interpretable score [27].

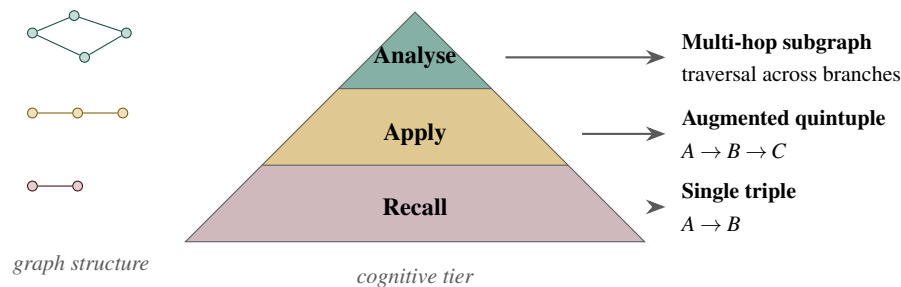


Figure 5: Conceptual schematic mapping cognitive tiers onto knowledge-graph traversal depth, after the construction in [27]. Graph topology also supplies plausible distractors: sampling sibling nodes from the same structural branch forces fine-grained conceptual discrimination rather than elimination of nonsense options.

5.3 Prompting versus Calibration: The Honest Distinction

The distinction that protects institutional credibility

The reality. The system's three-level classification relies on API prompt constraints and operational verbs. Prompt constraints approximate task complexity; they do *not* constitute psychometrically calibrated difficulty. True item response theory calibration requires fitting items against real student response telemetry over time [27, 14].

The defensible claim. The platform provides a transparent, structurally scaffolded pipeline that operationalises distinct cognitive actions — separating single-fact recall from procedural application and two-concept relational analysis. That is a real design property. Calling it “calibrated difficulty” would not be.

The pilot cohort is the route from the first statement to the second: response telemetry across a real module is precisely the input an IRT calibration requires.

6 Social Dynamics, Motivational Architecture and Psychological Safety

The platform's social layer — cohort feeds, aggregate analytics, gamified progress — sits at the intersection of self-determination theory and social-evaluative anxiety [28, 29].

6.1 The Pitfalls of Public Cohort Feeds

Public academic feeds are contested in the learning sciences. Intended to foster collaboration, they frequently function as arenas of social comparison and evaluation threat [1, 29].

Struggling and low-efficacy students consistently avoid posting in visible course environments, for fear of exposing academic deficits to peers and instructors [1, 2]. In professional or regulated programmes such as Medicine or Engineering, the concern that academic uncertainty might be noticed by faculty or professional bodies suppresses peer-to-peer engagement further [1, 4]. Public feeds therefore risk being monopolised by high performers, silencing exactly the avoidant students the platform exists to reach [1, 29].

6.2 De-Stigmatising Struggle Through Aggregate Insight

In contrast, the platform's **cohort insights** feature — displaying the aggregate proportion of students who answered a lecture quiz incorrectly, gated behind a minimum response threshold — is supported by social normative theory [30].

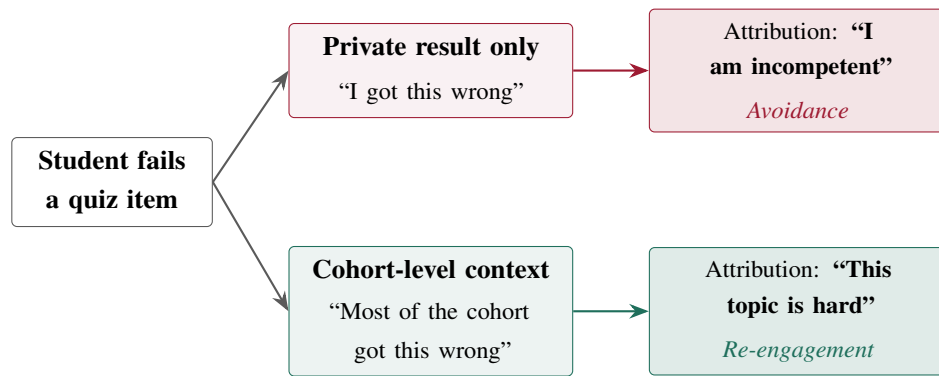


Figure 6: Conceptual schematic (not empirical data). Aggregate normative feedback redirects causal attribution from stable internal deficit to task difficulty, which is the mechanism by which it protects self-efficacy [30].

When a struggling student sees that a substantial share of the cohort also stumbled, internal deficit attribution is neutralised: the difficulty is read as intrinsic to the subject rather than personal inadequacy [1, 30]. By enforcing aggregation thresholds and omitting individual comparison entirely — never displaying “you are below average” — the feature supplies actionable feedback while preserving psychological safety [28, 30].

6.3 Autonomy in Gamification Architecture

Extrinsic gamification — points, streaks, daily goals, badges — carries documented motivational risk in higher education [28, 31]. Streaks can establish initial habits, but also introduce compulsive anxiety, displace intrinsic curiosity with point-seeking, and cause disengagement when broken [28, 31].

The universal rewards toggle, which disables streaks, points and badges *by default*, aligns with self-determination theory’s autonomy requirement [28]. Giving students explicit agency over whether they engage with gamified feedback prevents anxiety-inducing mechanics from harming neurodivergent or performance-anxious learners [28, 31]. Restricting leaderboards to an opt-in, pseudonymous format ensures competitive mechanics never expose vulnerable learners to scrutiny [1, 31].

Constraint

Defaulting rewards to *off* is defensible under self-determination theory but reduces measured engagement, and engagement is the metric a pilot is judged on. This is a genuine tension between what the literature supports and what the evaluation rewards. It should be declared in the pilot’s methodology rather than discovered in its results.

7 Synthesis: The Platform Against the Evidence

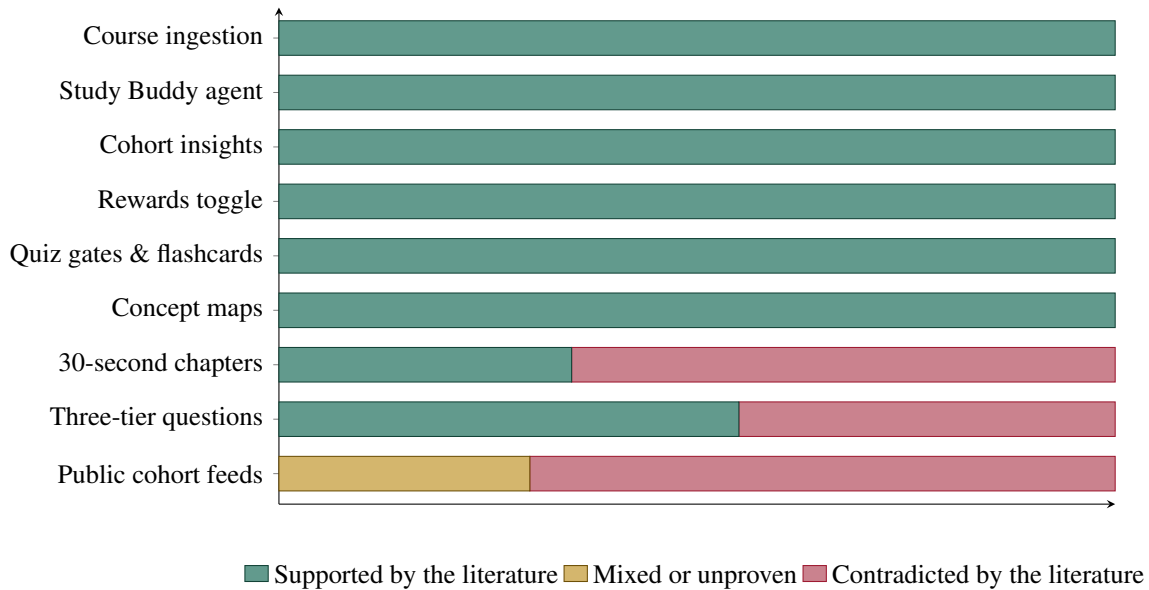


Figure 7: Authors' qualitative assessment, not measured data. Each feature is scored against the evidence reviewed above; bar proportions express the authors' judgement of how much of each feature's rationale survives scrutiny, and carry no statistical meaning. The two red-dominant rows are the platform's real exposure in an academic review.

Table 2: Feature-by-feature evaluation against the empirical literature.

Feature	Empirical alignment	Contradiction or constraint
Course ingestion	Supported. Eliminates administrative friction, dismantling the primary barrier to help-seeking [4, 9].	Relies on student-initiated upload; true equity requires institutional-level sync [4].
Study Buddy agent	Supported. Removes social-evaluative anxiety, giving a non-judgemental space for foundational questions [8, 10].	Must be strictly held to Socratic hinting or it reproduces executive offloading [11, 14].
Quiz gates & flashcards	Supported. Leverages the testing effect to consolidate memory and dismantle fluency illusions [13, 15, 16].	No cross-session spacing scheduler ; within-session testing yields diminishing returns [16].
30-second chapters	Partial. Supports re-entry after attentional lapses and respects working-memory limits [18, 19].	Arbitrary slicing lacks empirical support and risks severing the conceptual narrative [20, 22, 23].
Concept maps	Supported. Functions as an advance organiser anchoring topics in a domain hierarchy [23, 24].	Must be linked to question generation and segment browsing to preserve coherence [23, 27].
Three-tier questions	Supported. Enforces multi-concept synthesis at the Analyse tier, targeting Bloom's upper levels [25, 26].	Prompt-based tiers are not psychometrically calibrated difficulty [14, 27].
Cohort insights	Supported. Depersonalises failure through normative benchmarking, protecting self-efficacy [1, 30].	Requires minimum cohort density to guarantee privacy and reliability [30].
Public cohort feeds	Unproven. Contested; introduces acute social comparison and evaluation risk [1, 29].	Anxious students avoid posting , leaving forums dominated by high performers [2, 29].
Rewards toggle	Supported. Preserves autonomy under self-determination theory by making gamification opt-in [28, 31].	Gamification must stay secondary to conceptual feedback to avoid undermining intrinsic motivation [28].

7.1 What Follows

The platform operationalises sound cognitive mechanics: retrieval gating, friction removal in support delivery, structural mapping, and private normative feedback. Three positions require change rather than defence.

1. **Segment length should become an empirical question.** The thirty-second figure has no support. The pilot can test it directly by varying segment length across modules and measuring downstream quiz performance.
2. **Cross-session spacing should be built.** It is the single largest gap between what the system does and what the literature says works, and it is a scheduling problem rather than a research problem.
3. **Difficulty claims should stay descriptive until telemetry exists.** “Structurally scaffolded tiers” is defensible today. “Calibrated difficulty” becomes defensible only after item response data from a real cohort.

Naming these three openly is what distinguishes an evidence-based position from a marketing document, and it is the posture most likely to survive contact with the academics whose endorsement the platform requires.

A Citation Audit

Why this appendix exists

This appendix records what has and has not been verified against primary sources. A fabricated identifier causes more damage to a paper than any weakness in its argument, so the status of every post-2020 reference is stated explicitly rather than assumed. Three references remain unconfirmed at the time of publication and are marked as such in the bibliography. Corrections are welcome at the address on gradarina.com.

A.1 Corrections applied to the source draft

Table 3: Errors found and fixed during preparation.

Reference	Issue and resolution
Kessels & Steinmayr (2013)	The draft placed this in <i>Adults Learning</i> , 24(7), 234–247. Incorrect. Verified against ERIC, ScienceDirect and multiple citing works: it appeared in <i>Learning and Individual Differences</i> , 23, 234–240, DOI 10.1016/j.lindif.2012.09.013. Corrected.
“Information Editorial Review (2026)”	The draft cited a review of 41 classroom studies in <i>Information</i> , 17(3), 299, attributed to an editorial body rather than named authors. No such article could be located, and the attribution is not a valid author. Removed. The claim it supported — that fluent chatbot explanations inflate confidence without improving comprehension — is retained but now rests on Koriat & Bjork and Fan et al.
Şakiroğlu, Güvenir & Kaya (2026)	Verified as arXiv:2604.10748, submitted 12 April 2026. Upgraded to the peer-reviewed version: <i>Machine Learning and Knowledge Extraction</i> , 8(5), 137, DOI 10.3390/make8050137. Citing the journal version is materially stronger before an academic audience than citing a preprint.
Bjork & Bjork (2011)	The draft gave volume and issue numbers for what is a book chapter, not a journal article. Restructured as a chapter citation.

A.2 Still requiring verification

The following post-2020 references were carried over from the source draft and have **not** been checked against the publisher’s own record. Each should be opened on the journal page and the volume, issue, page range and DOI confirmed before external circulation:

- Kummer et al. (2024), ICIS proceedings — conference papers are the easiest category to mis-cite; confirm the proceedings year and paper number.
- Fan et al. (2024), *British Journal of Educational Technology* — confirm volume, issue and page range.
- Yan et al. (2024), *British Journal of Educational Technology* — confirm volume, issue and page range.

The pre-2010 references are long-established works in cognitive and educational psychology and carry low fabrication risk, but their page ranges have not been individually confirmed.

A.3 Sections where the evidence is thinnest

- **Section 4 (segmenting).** The argument against thirty-second slicing is well supported. The argument that concept maps *mitigate* fragmentation is an inference from advance-organisier research, not a direct finding about short-form video. It is presented as such.
- **Section 1 (ingestion).** The literature establishes that friction and stigma suppress help-seeking. It does not establish that removing them improves attainment via this mechanism. That is the pilot's job.
- **Figure 7.** Proportions express the authors' qualitative judgement. They are not measurements and must not be presented as such.

References

- [1] Karabenick, S. A., & Newman, R. S. (Eds.). (2006). *Help Seeking in Academic Settings: Goals, Groups, and Contexts*. Routledge.
- [2] Butler, R. (1998). Determinants of help seeking: Relations with competence, goals, and perceived classroom environment. *Journal of Educational Psychology*, 90(4), 630–643.
- [3] Karabenick, S. A., & Knapp, J. R. (1991). Relationship of academic help seeking to the approach and avoidance dimensions of achievement motivation. *Journal of Educational Psychology*, 83(2), 221–230.
- [4] Tinto, V. (2012). *Completing College: Rethinking Institutional Action*. University of Chicago Press.
- [5] Ryan, A. M., & Pintrich, P. R. (1997). “Should I ask for help?” The role of motivation and attitudes in adolescents’ help seeking in math class. *Journal of Educational Psychology*, 89(2), 329–341.
- [6] Addis, M. E., & Mahalik, J. R. (2003). Men, masculinity, and the contexts of help seeking. *American Psychologist*, 58(1), 5–14.
- [7] Kessels, U., & Steinmayr, R. (2013). Macho-man in school: Toward the role of gender role self-concepts and help seeking in school performance. *Learning and Individual Differences*, 23, 234–240. DOI:10.1016/j.lindif.2012.09.013
- [8] Aleven, V., Roll, I., McLaren, B. M., & Koedinger, K. R. (2016). Help helps, but only so much: Research on help seeking with intelligent tutoring systems. *International Journal of Artificial Intelligence in Education*, 26(1), 205–223.
- [9] Earl, W. R. (1988). Intrusive advising of freshmen in academic difficulty. *NACADA Journal*, 8(2), 27–33.
- [10] Kummer, T.-F., et al. (2024). Chatbots mitigate help-seeking avoidance: Evidence from underrepresented and introverted students in higher education. *Proceedings of the International Conference on Information Systems (ICIS 2024)*. [Unverified — see Appendix A]
- [11] Fan, Y., et al. (2024). Generative artificial intelligence and metacognitive laziness: Investigating the impacts of conversational agents on higher education learners’ self-regulated learning. *British Journal of Educational Technology*. [Volume and pages unverified — see Appendix A]
- [12] Koriath, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2), 187–194.
- [13] Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher et al. (Eds.), *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*. Worth Publishers.
- [14] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Hatzia Apostolou, T., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*. [Volume and pages unverified — see Appendix A]
- [15] Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181–210.
- [16] 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.

- [17] Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, 1(1), 69–84.
- [18] Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292.
- [19] Mayer, R. E. (2021). *Multimedia Learning* (3rd ed.). Cambridge University Press.
- [20] Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138.
- [21] Rey, G. D., Beege, M., Nebel, S., Wirzberger, M., Schmitt, M. E., & Schneider, S. (2019). A meta-analysis of the segmenting effect. *Educational Research Review*, 28, 100299.
- [22] Bradbury, N. A. (2016). Attention span during lectures: 8 seconds, 10 minutes, or more? *Advances in Physiology Education*, 40(4), 509–513.
- [23] Kalyuga, S. (2014). The expertise reversal principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd ed.). Cambridge University Press.
- [24] Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How Learning Works: Seven Research-Based Principles for Smart Teaching*. Jossey-Bass.
- [25] Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.
- [26] Agarwal, P. K. (2019). Retrieval practice and Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, 111(2), 189–209.
- [27] Şakiroğlu, M. C., Güvenir, H. A., & Kaya, K. (2026). Generating multiple-choice knowledge questions with interpretable difficulty estimation using knowledge graphs and large language models. *Machine Learning and Knowledge Extraction*, 8(5), 137. DOI:10.3390/make8050137
- [28] Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78.
- [29] Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology*, 96(3), 569–581.
- [30] Cialdini, R. B. (2007). Descriptive social norms as underappreciated sources of social control. *Psychometrika*, 72(2), 263–268.
- [31] Mekler, E. D., Brühlmann, F., Tuch, A. N., & Opwis, K. (2017). Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 71, 525–534.