Enhancing Lecture Comprehension: A Pilot Study of StashTag, a Real-Time Confusion Tracking and AI Remediation Tool

Soo, Wai Yam Minnie [1], Poon Tsz Hang [2] [1]Department of Physics, Hong Kong University of Science and Technology [2]Department of Computer Science and Engineering, Hong Kong University of Science and Technology

***Abstract*—The transition from high school to university-level STEM education presents challenges due to accelerated lecture pace, often leading to comprehension gaps that accumulate during instruction. This paper presents StashTag, a novel educational technology tool designed to combat "confusion accumulation" through real-time student feedback and AI-powered remediation. During a one-semester pilot across physics and mathematics courses (N=487 students), StashTag allowed students to anonymously timestamp confusing moments via a "confusion button." These timestamps generated personalized AI summaries and review questions post-lecture, while providing instructors with analytics on class-wide confusion patterns. Results from two-stage deployment show that engaged users rated the confusion button highly for usefulness (M=4.07/5) and the AI summaries for quality (M=4.17/5). However, longitudinal data revealed engagement decay due to usability barriers, metacognitive gaps in confusion identification and reporting, and workflow friction for instructors. Our findings highlight that while AI-enhanced tools show promise for personalized learning support, their sustained adoption requires designs that minimize friction, complete the confusion-resolution loop, and support metacognitive habit formation.**

***Index Terms*—Educational Technology, Artificial Intelligence in Education, Learning Analytics, STEM Education, Confusion Tracking, Real-time Feedback**

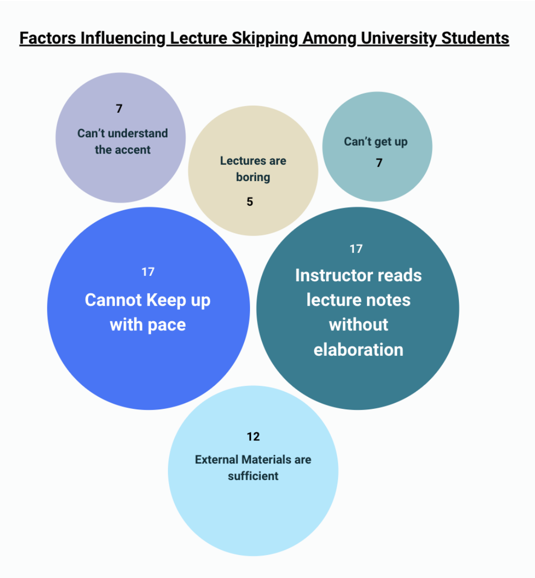


Fig. 1: Clustered analysis of open-ended responses from 90 university student participants, collected via an Instagram poll (showing the top clustered responses).

I. Introduction

The transition to university-level education, particularly in STEM (Science, Technology, Engineering, and Mathematics) disciplines, presents significant pedagogical challenges especially for first-year students. A primary obstacle is the accelerated pace of instruction compared to secondary schooling, which can quickly lead to comprehension gaps during live lectures. Preliminary surveys conducted confirm students' widespread perception that teaching speed is a barrier, correlating with declining voluntary lecture attendance (see Figure 1).

This pace-driven learning environment exacerbates a critical metacognitive problem: when a student becomes confused on a foundational concept during a fast-moving lecture, that moment of confusion can cascade, obstructing understanding of all subsequent materials. Students often struggle to both identify the precise source of their misunderstanding and formulate a clear question to resolve it, leading to what we term "confusion accumulation".

To address this challenge, we designed and developed StashTag, a novel educational technology tool. StashTag aims to disrupt the cycle of confusion accumulation by providing a low-friction, real-time mechanism for students to anonymously flag moments of confusion without disrupting the lecture flow. From the student's perspective, the system functions as a digital "SOS button"—a single click on a prominent interface element logs a timestamp. Post-lecture, artificial intelligence (AI) processes these timestamps, generating personalized "Confusion Summaries" that review and explain the lecture content from the flagged moments.

Simultaneously, StashTag serves as a learning analytics dashboard for instructors. By aggregating anonymized student confusion signals, the system visually identifies segments of the lecture that generated the highest collective confusion. This allows educators to review challenging topics, adjust future instruction, and, optionally, receive real-time notifications to address emerging misunderstandings immediately.

This paper presents a pilot study of StashTag's implementation across multiple physics and mathematics courses. We detail the system's design philosophy, its two-stage iterative development informed by user feedback, and an analysis of its impact on student engagement and perceived utility. Our findings contribute to the growing field of AI-enabled learning interventions, demonstrating a practical tool for bridging the gap between lecture delivery and student comprehension in large STEM classrooms.

II. Background and Development Trajectory of StashTag

StashTag responds to a widespread educational challenge known as "confusion accumulation," where students' unresolved misunderstandings during lectures progressively undermine their ability to achieve intended learning outcomes. This breakdown in comprehension ultimately manifests in poor academic outcomes, including consistently low average exam scores (often 50% or below) in physics, mathematics, and engineering courses (See TableI). Student commentary on USTSpace, the institution's official course evaluation platform, corroborates this pattern.

TABLE I: Historical Final Exam mean scores by course and semester.

Course & Semester	Final Exam Mean Score (%)
PHYS1112 Fall 2022	47.5
PHYS1112 Fall 2024	42.2
PHYS1112 Spring 2025	47.5
PHYS3033 Fall 2024	30.2
MATH1014 L1 Spring 2023	40.0
MATH2121 Fall 2023	48.3
MATH3423 Fall 2023	39.0
MATH3423 Fall 2025	28.3

. Professors report receiving negative student feedback through formal course evaluations (SFQ) at the semester's end, a delayed indication that student confusion was not identified and addressed in real time. Early-stage market research—conducted via Instagram polls, direct surveys, and in-depth interviews with over 100 students across two semesters (EMIA2020, 4900C/D)—revealed a fundamental disconnect in the lecture experience. This research crystallized into a clear Pain Point and Point of View (POV):

- **Student Pain Point:** The accelerated pace of university lectures creates comprehension gaps that compound during class time. Students struggle to identify and articulate these gaps during class, leading to passive disengagement and ineffective post-lecture review.
- **Instructor Pain Point:** Professors deliver content with limited immediate feedback (e.g. questions from students in-class or post-class) on student understanding, making it difficult to adapt pacing or clarify concepts before confusion cascades.
- **Core POV:** To bridge this disconnect, a tool must provide a frictionless way for students to signal confusion in the moment and receive automated, personalized remediation after class, while giving instructors actionable analytics on class-wide comprehension.

A. Evolution from Proof-of-Concept to Scalable System

The development followed a staged, lean methodology. A minimal viable product (MVP) was piloted in Spring 2025 (PHYS1114, PHYS1001), validating core user behavior (the "confusion button") through a semi-automated, labor-intensive process across seven lectures. The only automated feature was the button which marked the timestamps, but summaries were sent manually to the students. However, the feedback from students was positive and this proof-of-concept confirmed student willingness to engage with the tool and the value of the confusion data generated.

The current study represents the critical transition to a scalable, automated platform. In Fall 2025, we deployed a fully automated system, increasing pilot scale by over 8 times (from 7 to 59 lectures) and registering 487 students. This technical leap enabled reliable data collection on engagement and provided the foundation for robust feature development.

B. Validation Through Strategic Stakeholder Engagement

Beyond academic piloting, significant effort has been dedicated to validating the tool's market potential and securing growth pathways, demonstrating traction to institutional stakeholders and potential partners.

Initiative	Outcome / Indicator of Traction
Pitching to HKUST School of Science	Led to a dedicated Zoom session with 9 professors and the Center of Education Innovation (CEI), generating formal pilot interest.
Feature Publication in CEI Newsletter	Full interview and condensed promo video commissioned by CEI and displayed campus-wide, signaling institutional endorsement.
Consultation with HK Education Bureau	Active interest from the Course Development Department for potential application in secondary schools, indicating market expansion viability.
Partnerships with Secondary Schools	Two pilot agreements secured with La Salle College (Feb 2026) and Ti-I College (Jan 2026), validating the tool's applicability across educational levels.
Application to Cyberport Creative Microfund	Formal business plan and pitch developed; feedback received will refine future investment readiness.

TABLE II: Strategic Stakeholder Engagement and Traction Indicators

C. Structured Feedback and Iterative Design

A dual-channel feedback mechanism ensures continuous product refinement:

- **User Feedback:** Weekly syncs with a participating instructor to discuss system performance and pedagogical integration and two large-scale, incentivized student surveys (n≈50 each) provide direct input on usability and pedagogical impact.
- **Market Feedback:** Engagements with HKUST Center of Education Innovation (CEI), the Hong Kong Education Bureau and secondary schools validate the business model and expansion strategy.

III. System Architecture and Development

StashTag's system consists of two main parts: the front-facing student and professor pages, and the web server used for serving the frontend and processing collected data.

A. Infrastructure

StashTag uses Express.js and Node.js as the main framework for the web server. The server code is hosted in a private GitHub repository, which is connected to Google Cloud App Engine for hosting. Data collected from user testing is stored on a MongoDB Atlas database rather than on the persistent storage of the App Engine instance, as files within the storage are not persisted across server updates.

B. System Architecture

Figure 2 provides a high-level class diagram describing the system architecture.

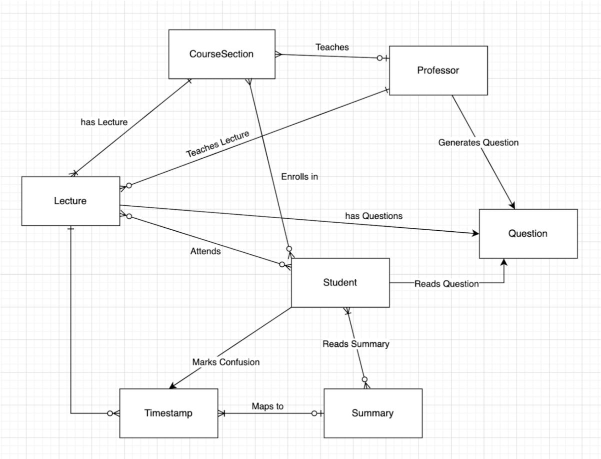


Fig. 2: High-level system architecture of StashTag

1) Lecture Transcription: During lectures, the system receives lecture audio, which is transcribed into text with corresponding start and end timestamps for each segment. For tests conducted during Fall 2025, Assembly AI's Slam-1 model was used for transcription. This will be migrated to ElevenLabs' speech-to-text model Scribe v1 for its multilingual support (99 languages including Cantonese), which will be required for expansion to local secondary schools.

Note: The audio used for transcription is immediately deleted from the App Engine instance after transcription completes.

2) Confusion Mapping and Summary Generation: Snippets of transcribed text are selected based on timestamps collected during the corresponding lecture by filtering for text within a 5-minute interval enclosing each timestamp. These snippets serve as context for DeepSeek v3.2 (deepseek-chat) to generate confusion summaries. Timestamps occurring at similar times map to the same snippet, so only one summary is generated per unique interval.

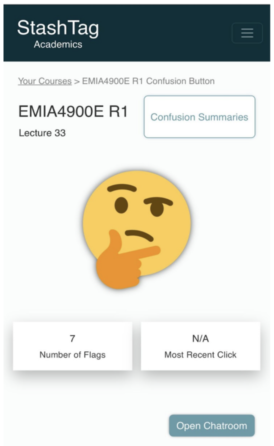
A system prompt enforcing formatting rules and requesting topic lists for each summary is included in the AJAX request to DeepSeek's services. Further improvements were made by incorporating ChatPhys, a RAG-LLM developed by HKUST alumni for physics courses. Generated summaries from DeepSeek are passed to ChatPhys' RAG model, which provides relevant images and lecture notes from PHYS1112 (support for additional courses has since been added).

C. Future System Improvements

1) Automated Email Notifications: Currently, email notifications (for class reminders, summary generation alerts, and pre/post-class questions) are sent manually. Implementing automated notifications will streamline this process and accommodate expected increases in testing volume.

2) Mobile Application: Existing frameworks such as Flutter or React.js enable multiplatform development (web, iOS, Android, macOS, Linux, Windows) and support technologies like mobile push notifications and screen wake locks, enhancing user experience and accessibility.

3) UI/UX Improvements: Survey feedback identified various interface issues causing navigation confusion and suggested modernizing the system's user interface. Figure 4 shows a draft interface incorporating some of these suggested changes.


StashTag
Academics

Your Courses > EMIA4900E R1 Confusion Button

EMIA4900E R1
Lecture 33

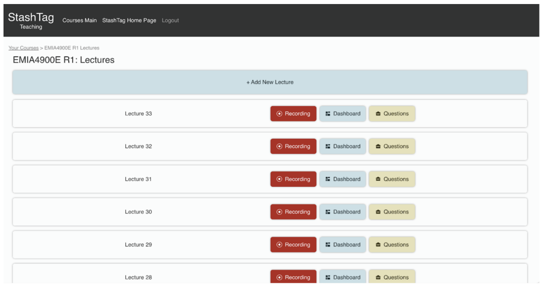
Confusion Summaries

7
Number of Flags

N/A
Most Recent Click

Open Chatroom

Fig. 3: Draft user interface (Confusion Button Page) incorporating suggested UX improvements


StashTag
Teaching
Courses Main StashTag Home Page Logout

Your Courses > EMIA4900E R1 Lectures

EMIA4900E R1: Lectures

+ Add New Lecture

Lecture 33 Recording Dashboard Questions

Lecture 32 Recording Dashboard Questions

Lecture 31 Recording Dashboard Questions

Lecture 30 Recording Dashboard Questions

Lecture 29 Recording Dashboard Questions

Lecture 28 Recording Dashboard Questions

Fig. 4: Draft user interface (Professors' lectures page) incorporating suggested UX improvements

IV. Pilot Study: Focused Analysis on an undergraduate level Physics course – PHYS1112

A. Study Context and Participants

A pilot study of StashTag was conducted across four course sections (PHYS1112 L1, L3, L6; MATH3423 L1) at Hong Kong University of Science and Technology. The study involved three distinct instructors and a total of 487 registered undergraduate students. Instructors were recruited via direct contact and provided with varying onboarding methods including written manuals, video manuals and face-to-face verbal training.

TABLE III: Courses in which StashTag was Piloted

Courses	Lectures	Students	Notes
PHYS1112 L1	20	139	
PHYS1112 L3	18	122	
PHYS1112 L6	6	110	
MATH3423 L1	8	60	
PHYS1112 L2	3	10	Not included in analysis
PHYS1114 L1	1	7	Not included in analysis
PHYS1114 L2	3	12	Not included in analysis

B. Tool Deployment and In-Class Protocol

The core functionality of StashTag required instructors to initiate a session via a dedicated web portal (https://stashtag.df.r.appspot.com/courses). Upon logging in (requiring re-authentication every 45 days of inactivity), instructors navigated to a recording interface to begin a lecture session. The system then generated a unique, session-specific QR code for student access.

At the start of each lecture, instructors displayed the QR code and delivered a standardized introduction explaining the tool's purpose and functionality. As evidenced by instructor testimonials, these introductions framed StashTag as a student-developed learning enhancement tool designed for math-intensive courses, where complex derivations can lead to accumulating confusion. Students were encouraged to press a prominent "confusion button" on their interface whenever they experienced misunderstanding during the lecture.

C. Data Capture and Real-Time Feedback

When a student activates the confusion button, the system records a precise timestamp. These anonymized timestamps are aggregated and displayed immediately after the lecture ends, on an instructor Dashboard as a temporal bar chart, plotting frequency of confusion signals against lecture time. This visualization allows instructors to identify "confusion peaks" after each session (see Figure 5). This allows instructors to provide further clarification or elaboration in the following class. Clicking a bar initially displayed the corresponding segment of the automated lecture transcript (Stage 1), and was later enhanced to also show AI-generated student summaries (Stage 2) (see Figure 6).

D. Post-Lecture AI-Powered Remediation

Following the lecture, the system processes each recorded timestamp. It extracts a contextual audio clip (spanning from 4 minutes before to 1 minute after the timestamp) which is transcribed. This transcript snippet is fed into the DeepSeek-chat LLM via a specialized prompt engineered to generate a Confusion Summary. This summary, aimed at clarifying the potentially confusing concept, is delivered to the respective student via a personal "Confusion Summaries" page approximately 30 minutes post-lecture (see Figure 7).

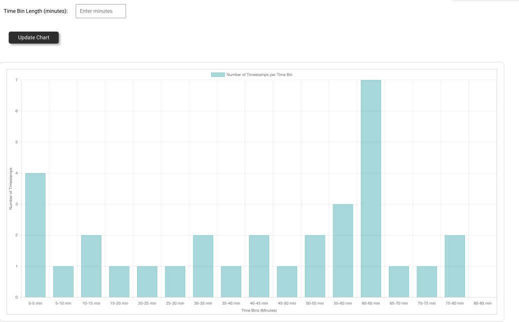
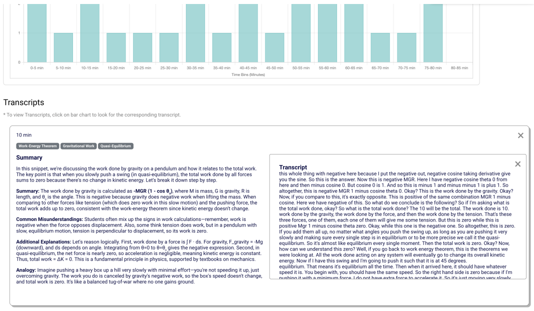


Fig. 5: Instructor dashboard displaying the number of confusion presses (timestamps collected) plotted against the lecture time (5 min intervals)



Transcripts

* To view Transcripts, click on bar chart to look for the corresponding transcript.

10 min

Work-Energy Theorem | Gravitational Work | Quasi-Equilibrium

Summary

In this snippet, we're discussing the work done by gravity on a pendulum and how it relates to the total work. The key point is that when you slowly push a swing (in quasi-equilibrium), the total work done by all forces sums to zero because there's no change in kinetic energy. Let's break it down step by step.

Summary: The work done by gravity is calculated as **-MGR (1 - cos θ_0)**, where M is mass, G is gravity, R is length, and θ_0 is the angle. This is negative because gravity does negative work when lifting the mass. When comparing to other forces like tension (which does zero work in this slow motion) and the pushing force, the total work adds up to zero, consistent with the work-energy theorem since kinetic energy doesn't change.

Common Misunderstandings: Students often mix up the signs in work calculations—remember, work is negative when the force opposes displacement. Also, some think tension does work, but in a pendulum with slow, equilibrium motion, tension is perpendicular to displacement, so its work is zero.

Additional Explanations: Let's reason logically. First, work done by a force is ∫ F · ds. For gravity, F_gravity = -Mg (downward), and ds depends on angle. Integrating from θ=0 to $\theta=\theta_0$ gives the negative expression. Second, in quasi-equilibrium, the net force is nearly zero, so acceleration is negligible, meaning kinetic energy is constant. Thus, total work = ΔK = 0. This is a fundamental principle in physics, supported by textbooks on mechanics.

Analogy: Imagine pushing a heavy box up a hill very slowly with minimal effort—you're not speeding it up, just overcoming gravity. The work you do is canceled by gravity's negative work, so the box's speed doesn't change, and total work is zero. It's like a balanced tug-of-war where no one gains ground.

Transcript

this whole thing with negative here because I put the negative out, negative cosine taking derivative give you the sine. So this is the answer. Now this is negative MGR. Here I have negative cosine theta 0 from here and then minus cosine 0. But cosine 0 is 1. And so this is minus 1 and minus minus 1 is plus 1. So altogether, this is negative MGR 1 minus cosine theta 0. Okay? This is the work done by the gravity. Okay? Now, if you compare to this, it's exactly opposite. This is positive of the same combination MGR 1 minus cosine. Here we have negative of this. So what do we conclude is the following? So if I'm asking what is the total work done, okay? So what is the total work done? The 10 will be the total. The work done is 10. work done by the gravity, the work done by the force, and then the work done by the tension. That's these three forces, one of them, each one of them will give me some tension. But this is zero while this is positive Mgr 1 minus cosine theta zero. Okay, while this one is the negative one. So altogether, this is zero. If you add them all up, no matter what angles you push the swing up, as long as you are pushing it very slowly and making sure every single step is in equilibrium or to be more precise we call it the quasi-equilibrium. So it's almost like equilibrium every single moment. Then the total work is zero. Okay? Now, how can we understand this zero? Well, if you go back to work energy theorem, this is the theorems we were looking at. All the work done acting on any system will eventually go to change its overall kinetic energy. Now if I have this swing and I'm going to push it such that it is at 45 degrees. equilibrium. That means it's equilibrium all the time. Then when it arrived here, it should have whatever speed it is. You begin with, you should have the same speed. So the right hand side is zero because if I'm pushing it with a minimum force, I do not have extra force to accelerate it. So it's just moving very slowly

Fig. 6: Students' Confusion summaries displayed to the instructor after clicking onto its corresponding bar.

E. Iterative Design and Feature Development

The pilot employed a design-based research approach, with significant refinements made between a first and second stage of testing based on user feedback.

1) Authentication Flow: The initial email-verification login posed a cross-device compatibility hurdle (e.g., scanning QR with a phone but verifying on a laptop). This was first mitigated with a manual "verify button" workaround and later replaced entirely with a more reliable password-based authentication system.

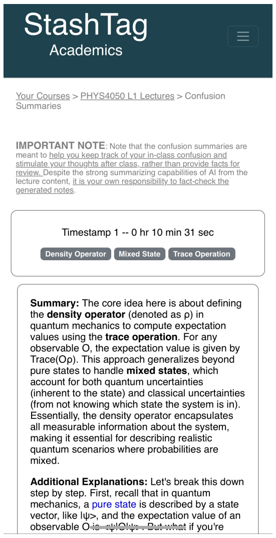
2) AI Prompt Engineering: The prompt for the DeepSeek-chat model was refined from an informal, friendly tone (Stage 1) to a structured template ensuring consistent output containing: a concise summary, supplemental explanations, and "food for thought" questions (Stage 2).

3) Enhanced Instructor Dashboard: Based on professor feedback, the dashboard was upgraded. Instead of solely displaying the lecture transcript for a confusion peak, it integrated the students' AI-generated confusion summaries, allowing instructors to compare their delivered content against student interpretations.

F. Advanced Feature Rollout in Stage 2

Building on the core functionality, several advanced features were deployed in the second testing stage:

1) **Keyword Tagging:** Each AI-generated confusion summary was automatically tagged with three relevant conceptual keywords.



Your Courses > PHYS4050 L1 Lectures > Confusion Summaries

IMPORTANT NOTE: Note that the confusion summaries are meant to help you keep track of your in-class confusion and stimulate your thoughts after class, rather than provide facts for review. Despite the strong summarizing capabilities of AI from the lecture content, it is your own responsibility to fact-check the generated notes.

Timestamp 1 -- 0 hr 10 min 31 sec

Summary: The core idea here is about defining the **density operator** (denoted as ρ) in quantum mechanics to compute expectation values using the **trace operation**. For any observable O, the expectation value is given by Trace(Oρ). This approach generalizes beyond pure states to handle **mixed states**, which account for both quantum uncertainties (inherent to the state) and classical uncertainties (from not knowing which state the system is in). Essentially, the density operator encapsulates all measurable information about the system, making it essential for describing realistic quantum scenarios where probabilities are mixed.

Additional Explanations: Let's break this down step by step. First, recall that in quantum mechanics, a pure state is described by a state vector, like |ψ>, and the expectation value of an observable O is <ψ|O|ψ>. But what if you're

Fig. 7: Confusion summaries: Summaries and additional explanations corresponding to students' confusions displayed to the students after lecture.

2) **RAG-LLM Integration:** A collaboration with ChatPhys enriched physics-related summaries by retrieving relevant excerpts and diagrams from official lecture notes.
3) **Keyword Search Function:** Students gained the ability to search across all their confusion summaries using conceptual keywords.
4) **Review Question Generation:** An additional AI module was implemented to generate multiple-choice review questions based on confusion summaries.

V. Results and Analysis

A. Stage 1: Initial User Feedback and Identification of Barriers

A post-deployment survey was administered to the 386 students registered during the initial pilot stage, yielding 52 usable responses (13.5% response rate). The tool received an overall mean rating of 3.85 (SD = 0.98) on a 5-point scale. The core AI-generated confusion summary feature was rated 3.59 (SD = 1.26), indicating moderate initial satisfaction. When asked about the summaries' impact, a combined 70.3% of respondents reported they clarified concepts either "a little" (48.1%) or "most" of the confusions (22.2%). Furthermore, 37.0% indicated the summaries aided in visualizing concepts more effectively.

The survey identified significant friction points hindering adoption. Specifically, 37.0% of respondents found the initial email-verification login process confusing. An additional comprehension gap was evident, as 7.4% reported waiting for summaries to arrive via email, despite the summaries being accessible only on a dedicated platform page. This suggests deficiencies in user guidance regarding the tool's post-lecture workflow.

Regarding in-class use, 47.0% of respondents reported not using the Confusion button because they "did not feel the need" to do so during lectures. When queried about potential future enhancements, students expressed strong interest in more interactive features. On a 4-point desirability scale (4 = most desirable), the following features received the highest rating (4/4) from a substantial portion of respondents: the ability to ask questions anonymously in class (50.0%), inclusion of textbook exercises aligned with confused concepts (42.3%), and functionality for follow-up dialogue with the StashTag AI (48.1%).

B. Stage 2: Enhanced Deployment and Refined Engagement

Following the implementation of design improvements, a second survey was administered (n=62, 12.9% response rate). Within this cohort, 75.8% (n=47) of respondents were aware of the core Confusion button feature. However, awareness of the linked remediation features showed a significant drop: only 43.5% (n=27) were aware of the AI-generated confusion summaries, and 32.3% (n=20) were aware of the review questions. This limitation reduces the reliability of Stage 1 feedback, as the analysis assumed all respondents had direct experience with the platform's features.

TABLE IV: Stage 2 Perceived Utility and Usability Ratings (M ± SD)

Feature / Dimension	n	Rating (1–5 Scale)
Confusion Button		
Convenience	47	4.02 ± 0.82
Usefulness	47	4.07 ± 0.88
Confusion Summaries		
Overall Quality	17	4.17 ± 0.85
Convenience	17	4.13 ± 0.72
Usefulness	17	4.29 ± 0.77
Summary Adjuncts		
Diagrams from Notes	17	4.00 ± 0.96
Reference Links	17	3.77 ± 1.09
System & Support		
User Interface (UI)	17	3.77 ± 0.90
Clarity of Instructions	17	3.94 ± 1.09
Communications/Onboarding	17	3.76 ± 1.03
Quality of Review Questions	17	4.25 ± 0.93
Review Questions		
Convenience	11	4.09 ± 1.04
Usefulness	11	4.18 ± 1.08

Survey data revealed a steep attrition in user engagement from awareness to sustained feature use: 77.4% used the Confusion button, but only 37.5% of those users proceeded to the summaries. Usage of advanced features was minimal: keyword search (8.3%) and review questions (25.0%). Among users of confusion summaries (n=13), engagement was typically brief: 38.5% spent less than one minute, 30.8% spent 1–5 minutes, 23.1% spent 6–10 minutes, and 7.7% spent 10–15 minutes reviewing the content. These findings indicate that students primarily use confusion summaries as a quick remediation tool, rather than for in-depth review or long-term revision.

Ratings from users who engaged with the core features were consistently positive (Table IV). Open-ended responses indicated that students valued the summaries for their ability to "condense material into key points for rapid review" and to help "revisit areas of uncertainty."

C. Longitudinal Engagement Patterns

Analysis of system log data from PHYS1112 L1 and L3 reveals significant temporal trends in user engagement, defined as the rate of Confusion button presses per lecture.

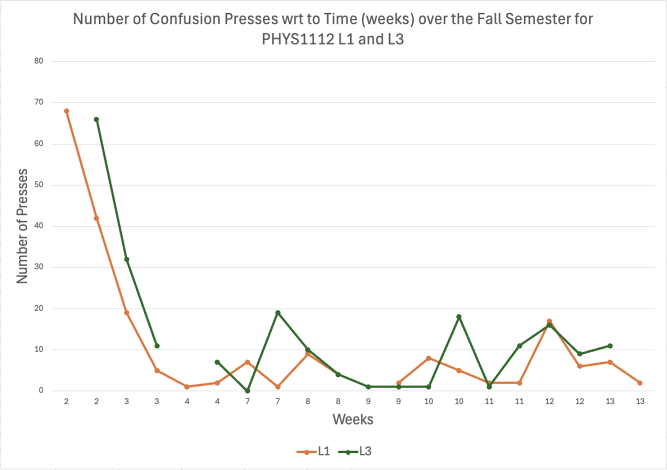


Fig. 8: Longitudinal engagement trends (Confusion button presses per lecture) for PHYS1112 sections L1 and L3. The first three minutes of each lecture were excluded to filter trial presses. Breaks in the graph indicates that there is no lecture.

Engagement demonstrated a characteristic pattern of high initial adoption followed by a steep decline (Fig. 14). Both sections exhibited peak usage in the first two lectures, a period coinciding with instructor introductions. However, engagement rapidly decayed in subsequent lectures due to usability barriers, natural attenuation of novelty, and a potential mismatch between student metacognitive habits and the tool's required in-the-moment reporting action.

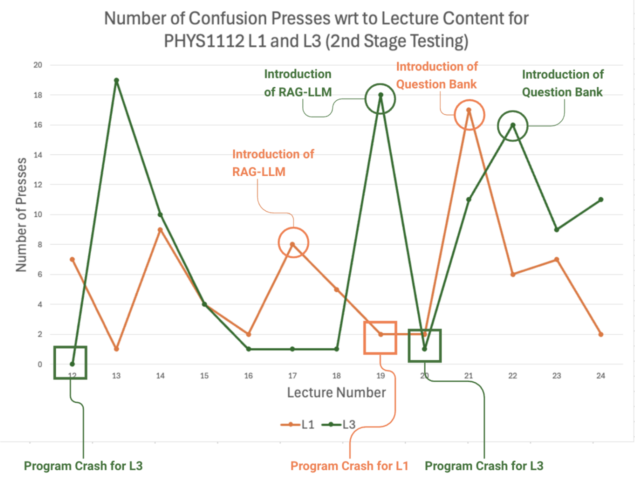


Fig. 9: Engagement trends based on lecture number (which corresponds to lecture content) of the 2nd Piloting Stage starting from lecture 12 (Confusion button presses per lecture) for PHYS1112 sections L1 and L3. Circles indicate introduction of a new feature, and squares indicate program crashes.

The longitudinal data also shows distinct engagement spikes occurring mid-semester, which correspond temporally with announcements of new system features. These spikes indicate that feature awareness, often prompted by instructor reminders, is a primary driver of renewed usage. A notable spike in L3 during Lecture 13 and 21 occurred without a new feature launch but immediately followed a general instructor reminder, suggesting that sustained engagement is highly dependent on periodic prompting.

System crashes are associated with a drop in engagement. However, when instructors subsequently announced that the system had been restored, usage frequently rebounded, often producing a spike. This pattern reinforces the interpretation that engagement is instructor-dependent, and students appear to lapse in usage absent reminders, suggesting limited habit formation and a tendency to forget or deprioritize the system without external cues.

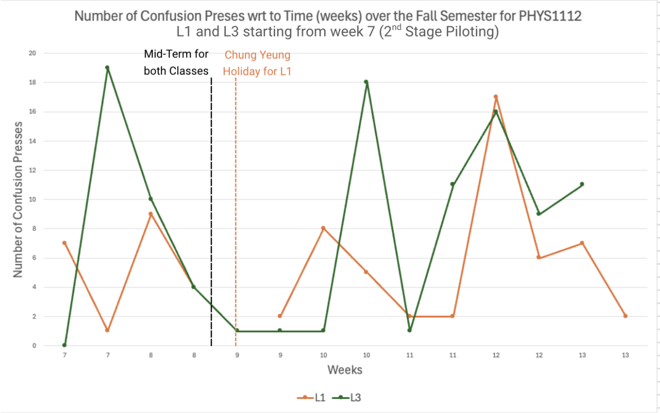


Fig. 10: Longitudinal engagement trends of the 2nd Piloting Stage starting from week 7 (Confusion button presses per lecture) for PHYS1112 sections L1 and L3. Dotted lines indicate special events during the semester, including mid-terms and holidays.

The observed decline in engagement may also be attributed to external academic cycles, particularly mid-term examinations. The downward trend following the post-reminder spike in Lectures 14 and 15—a pattern consistent across both sections—lends support to this interpretation.

We did not find a correlation between the number of presses in L1 and L3 when we grouped activity by lecture content ((Fig. 11) We had expected a positive relationship—if L3 had more presses, L1 would as well—because confusion should be driven by lecture content and produce similar peaks across both sections. Instead, the patterns appear irregular. This likely reflects section-specific events (e.g., program crashes, new feature rollouts, and the timing of instructor reminders) that introduce noise and reduce our ability to detect any content-driven relationship.

Readers may also be interested in whether the confusion presses expressed in the data can approximate the number of unique users during the pilot and thereby serve as a valid proxy for overall class engagement. Our analysis indicate that the total number of presses is moderately correlated with the count of unique users, suggesting that press frequency provides a reliable representation of aggregate engagement.

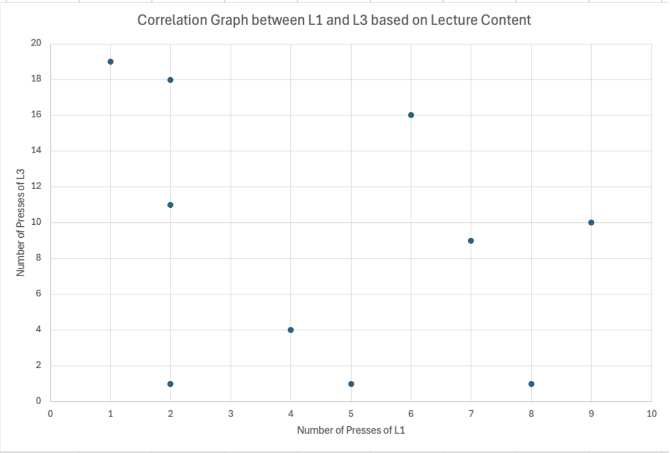


Fig. 11: Press activity for sections L1 and L3. No correlation is observed between sections when counts are grouped by lecture.

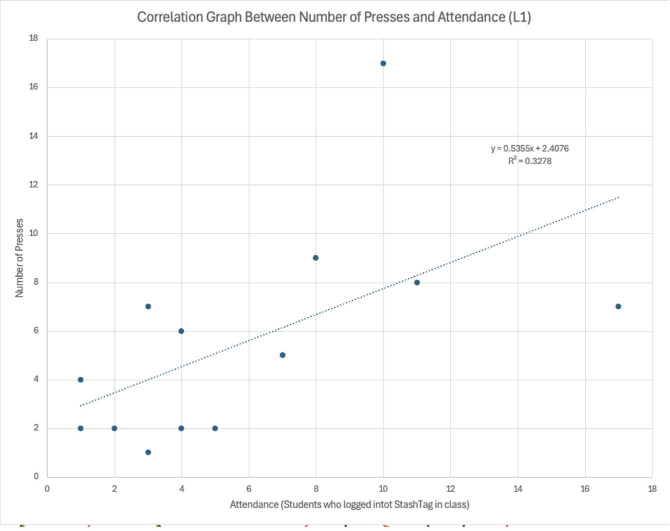


Fig. 12: Number of presses and attendance of PHYS1112 L1 over the entire semester.

D. Student Qualitative Feedback

Through analysis of open-ended survey comments, students found the tool helpful for quickly capturing confusion, getting concise summaries of key points, and using review questions to reinforce understanding. The Confusion button, when working smoothly, was particularly valued for its speed and simplicity. Common requests included clearer guidance (built-in help/FAQ), mobile interface simplification, improved reliability, better tracking of flagged confusions, and more accurate, tailored summaries (See Table V).

E. Instructor Feedback

Qualitative feedback from two instructors revealed favorable evaluations of the tool's outputs. One professor noted the AI-generated confusion summaries were of "good quality" with "additional and correct information" not covered in lecture. AI-generated review questions were also seen as "good quality" and a potential workload alleviator.

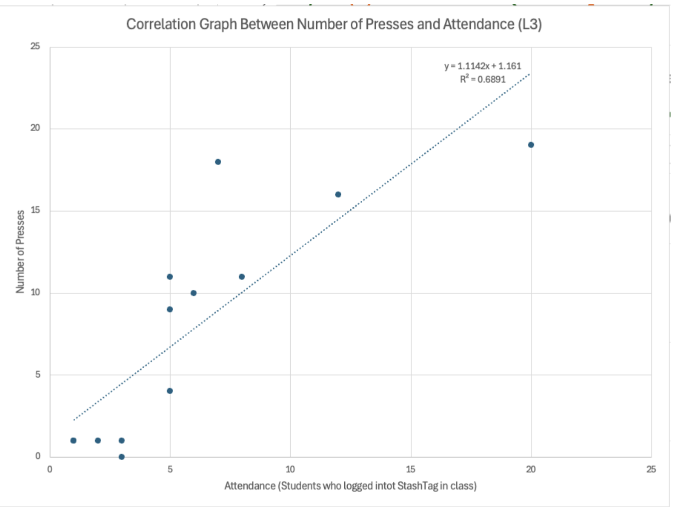


Fig. 13: Number of presses and attendance of PHYS1112 L3 over the entire semester.

TABLE V: Selected qualitative feedback from student surveys.

Category	Student Comment
Positive Feedback	"The summaries and the review questions help me to review knowledge and better understand them." "It concentrated a wide range of knowledge to a summary which help me can find the key point quickly." "Review questions are good because it clarifies the topic further." "Very innovative." "It is very good already that it helps me summarise the difficulties."
Appreciation & Interest	"Thank you for making this button for us! Quite interesting, but still needed to familiar with this function."
Technical & UI Issues	"The UI is not very good and the one time that we tried to use it in class it straight up didn't work." "Interface could be improved much better. I use it on mobile and it was kind of confusing. Like I didn't know what I press would trigger what." "And the confusion emoji was way too big that if I was just trying to scroll, I would accidentally press it."
Suggestions for Improvement	"A FAQ/help function can be included in the webpage so it's possible for us to figure it out myself, since the Instructor didn't really know how to use StashTag during class and it took a little bit of delay." "A little more detail on how to use all features." "It's hard for me to sidetrack on loading the page and use the confusion button when I'm already in confusion."

However, instructors cited practical barriers preventing consistent integration. A primary challenge was workflow integration—difficulty recalling and accessing the tool's URL immediately before class was cited as "too time consuming." The perception of added workload was another critical barrier, with one instructor emphasizing a desire to avoid "extra workload."

Instructors suggested that AI-generated review questions would serve as excellent preparatory material if displayed to students before lectures for self-testing. One instructor emphasized that "learning is a very self-initiated process," underscoring that the tool's effectiveness is contingent on student motivation.

VI. Discussion

A. Confusion Tracking as a Mechanism for Improving Instructional Efficiency

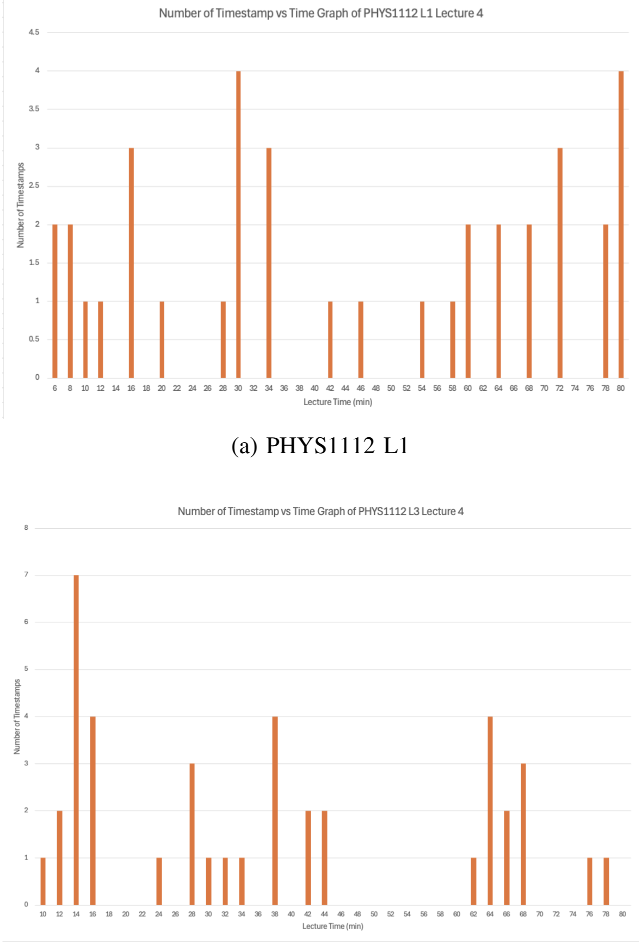
The high initial engagement observed across pilot sections (Lecture 3 and lecture 4) suggests that students readily identified with the core problem of confusion accumulation (Fig.14a,Fig.14b). The promise of systematically capturing these moments in real-time provided strong initial incentive. Analysis of unbiased click data reveals that students' confusion was distributed throughout lectures in both L1 and L3 sections, indicating that moments of misunderstanding are not isolated to specific topics but represent a pervasive challenge during instruction. This finding underscores the potential impact of StashTag's approach. The prominent confusion peaks identify specific topics that require additional instructor clarification, while the wide distribution of confusion points reflects the difficulty of addressing all student needs simultaneously during live instruction. StashTag addresses this challenge by delivering automated, personalized confusion summaries.

B. The Challenge of Sustaining User Incentive

Longitudinal engagement data reveals a common pattern in educational technology: initial curiosity alone is insufficient to sustain long-term use. Rather than relying solely on prompts to reinforce user habits, both student and instructor engagement depends on consistently delivering immediate, tangible value with minimal friction.

1) Student Incentive: Closing the Confusion-Resolution Loop: Students demonstrated clear recognition of the tool's potential, but incentive decayed when the perceived workflow was incomplete. The critical insight is that tracking confusion is only motivating if resolution reliably follows. The engagement spike following RAG-LLM integration and more sustained engagement after review questions suggest that incentive is maximized by a complete cognitive loop: identifying a knowledge gap → receiving targeted explanation → actively testing restored understanding.

2) Instructor Incentive: The Primacy of Minimizing Friction: Instructor feedback underscores that their incentive calculus is dominated by ease of integration and zero marginal workload. The professors' adoption challenges are explained by the compensatory relationship between motivation and ability in BJ Fogg's behavior model. While initially motivated, this motivation was quickly depleted by low ability—friction points like browser incompatibility made the simple act of starting the system unexpectedly difficult. Consequently, the effort required began to outweigh the perceived benefit, leading to attrition.



(a) PHYS1112 L1

(b) PHYS1112 L3

Fig. 14: In-lecture engagement patterns (confusion button presses for every 2 minutes) for representative lectures in (a) PHYS1112 L1 and (b) PHYS1112 L3. Both sections show concentrated activity immediately following the instructor's introduction, demonstrating consistent initial student receptivity to the tool's proposed utility.

C. Multi-Layered Barriers to Adoption

Beyond incentive structures, our findings reveal barriers at behavioral, cognitive, and systemic levels.

1) Behavioral and Cognitive Hurdles: A significant barrier is the metacognitive gap between experiencing and actively reporting confusion. The discrepancy between universal self-reported confusion and the portion who did not press the button suggests many students do not engage in real-time metacognitive monitoring. This points to a need for pedagogical scaffolding to normalize confusion-tagging as a positive learning act.

Furthermore, student feedback indicates a preference for low-effort, high-efficiency review. The brief time spent with summaries and higher valuation of diagrams over textual links indicate a desire for quick, visual clarification, creating a design tension between comprehensive explanations and cognitive load.

2) Systemic and Contextual Barriers: The steep feature awareness-attrition funnel (75.8% button awareness vs. 37.5% summary usage) highlights a critical systemic failure in communication and onboarding. This barrier is compounded by external academic rhythms, such as mid-term examinations, which redirect student attention and disrupt nascent usage habits.

D. Theoretical and Practical Implications

The findings reinforce that utility alone does not drive adoption in educational tools; success depends equally on frictionless integration, clear communication of value, and support for user metacognition. Practically, for tools like StashTag to thrive, development must focus on: (1) completing the learning loop, (2) designing for frictionless integration, and (3) embedding pedagogical support.

VII. Limitations and Future Plans

This semester, while deploying our product in real classroom settings, we encountered several limitations that impacted our findings and overall effectiveness. One significant limitation was the design of our questionnaires. Although we gathered genuine feedback from a substantial number of students, the questionnaires were not optimally structured for data analysis. The design of high-level questions and overly complex multiple-choice options hindered our ability to extract meaningful and quantifiable insights leading to a waste of feedback data. A more refined approach, such as using Likert scales in more questions, would improve our data collection and analysis in future iterations.

Additionally, the system crashes during presentations led to inconsistencies in our results. These crashes not only disrupted the flow of our demonstrations but also introduced spontaneity into the data collected. For instance, the lack of correlation between the number of presses in L1 and L3 suggests that external factors—such as program crashes and feature rollouts—created noise that masked potential content-driven relationships. As we proceed, it is crucial to establish standardized testing procedures to ensure our system functions reliably across diverse devices and settings.

Moving forward, we plan to refine our questionnaires to focus on specific data collection aimed at meaningful analysis. Moreover, we will prioritize enhancing the system's stability through rigorous testing and iterative improvements, enabling us to gather more accurate feedback and drive better design decisions. Finally, we aim to enhance user awareness of core features to ensure that feedback reflects a comprehensive understanding of our product, thus bolstering the reliability of our findings.

VIII. Conclusion

This pilot study demonstrates the significant potential of real-time confusion tracking paired with AI-powered remediation. The tool successfully validated its core premise: students actively engaged when provided with a clear, low-friction path from identifying confusion to receiving targeted explanations and practice. However, sustained adoption hinges on overcoming critical barriers—most notably, seamless integration into existing instructor workflows and designing for the metacognitive habits of students. Future iterations that prioritize frictionless access and complete learning feedback loops can transform StashTag from a promising prototype into a robust, scalable support for STEM education.

Acknowledgments

The authors would like to thank the participating students and instructors at HKUST, as well as the developers of the ChatPhys system for their collaboration.