# EchoMind: Augmenting Recorded Presentations with Context-Grounded Conversational Agents and Collective Viewer Signals


Haiyang Xu*
Department of Computer Science
Aarhus University
Aarhus, Denmark
202402604@post.au.dk

Yunhui Song*
Department of Computer Science
Aarhus University
Aarhus, Denmark
202402597@post.au.dk

Black Sun
Department of Computer Science
Aarhus University
Aarhus, Denmark
202403892@post.au.dk

Niklas Elmqvist
Department of Computer Science
Aarhus University
Aarhus, Denmark
elm@cs.au.dk

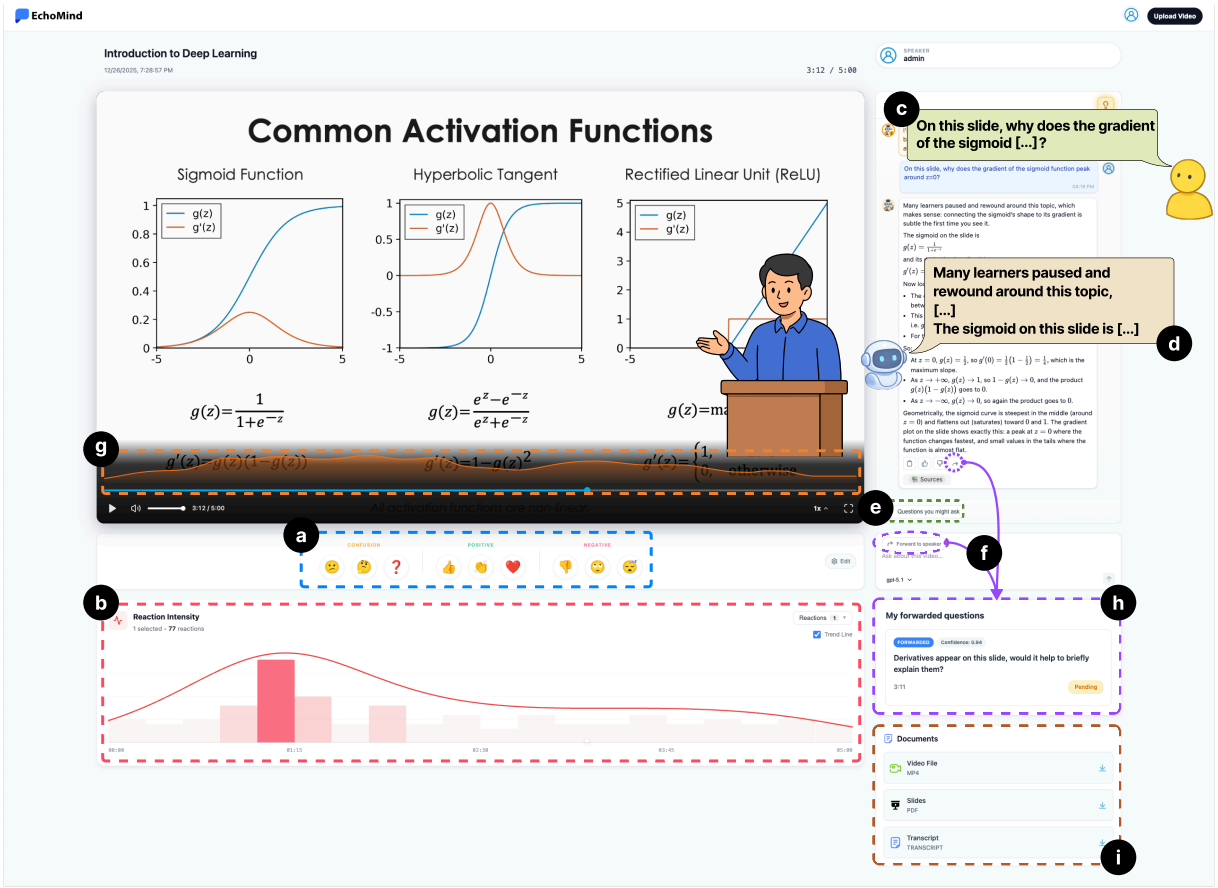



**Figure 1: EchoMind viewer interface. Viewers interact with the video through an emoji-based reaction interface (a), alongside natural video interactions such as pausing, replaying, and seeking, while aggregated audience reactions are visualized over time (b). During viewing, viewers can ask context-specific questions (c), which are answered by a conversational agent that leverages the current slide, video context, and collective audience behavior (d). When needed, viewers can escalate their questions using an escalation mechanism (f), either by directly forwarding them to the speaker or by forwarding them after AI-assisted clarification. When entering a question, viewers can use question input suggestions (e) to quickly formulate questions grounded in the current video context. These reactions, viewing behaviors, and questions jointly expand the system's knowledge base and are accumulated into a temporal learning signal curve (g). Escalated questions are collected and displayed in a panel (h), waiting for speaker's response. Materials, transcripts, and documents are shown and made available for download (i). Video source: video in Section 5.1.**


## Abstract

Recorded presentations have become pervasive in education, yet they often reduce knowledge transfer to a one-way broadcast: viewers cannot clarify their confusion or gauge whether others share their struggles. We present EchoMind, a system that augments recorded presentations with a conversational agent grounded in presentation materials and collective viewer signals (behavioral traces and affective reactions), alongside timeline-aligned community awareness visualizations. In a between-subjects study (N=75), participants watched a 5-minute recorded lecture under one of three conditions: video-only, video with a generic LLM chatbot, and video with EchoMind. EchoMind significantly improved the subjective learning experience across multiple dimensions, including confusion clarification, in-the-moment expression, and co-learning perception, without increasing perceived workload. The generic LLM condition did not significantly outperform video-only and received the lowest mean ratings on usability. These findings suggest that contextual grounding and community awareness, rather than the presence of conversational AI alone, make AI a meaningful support in asynchronous video learning.



## CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Collaborative and social computing systems and tools**; *Empirical studies in HCI*.

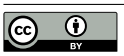
## Keywords

Asynchronous learning, conversational agents, recorded presentations, retrieval-augmented generation.

**ACM Reference Format:**

Haiyang Xu, Yunhui Song, Black Sun, and Niklas Elmqvist. 2026. EchoMind: Augmenting Recorded Presentations with Context-Grounded Conversational Agents and Collective Viewer Signals. In *Proceedings of the 14th Nordic Conference on Human-Computer Interaction (NordiCHI '26), October 03–07, 2026, Vaasa, Finland*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3829807.3829883


*Both authors contributed equally to this research.




This work is licensed under a Creative Commons Attribution 4.0 International License.


*NordiCHI '26, Vaasa, Finland*


© 2026 Copyright held by the owner/author(s).


ACM ISBN 979-8-4007-2373-5/26/10
https://doi.org/10.1145/3829807.3829883


## 1 Introduction

Recorded online presentations have become ubiquitous in academic and educational settings, particularly following the widespread adoption of remote instruction during the COVID-19 pandemic [2, 35]. However, as Moore's classic framework suggests, effective distance education involves multiple forms of interaction, including learner–content, learner–instructor, and learner–learner interaction [49]. Recorded presentations primarily support learner–content interaction, while offering limited opportunities for learners to ask instructors clarifying questions or benefit from interactions with other learners [3]. As a result, asynchronous viewers often lack timely support when confusion arises and may need to rely on replay, selective revisiting, or delayed follow-up channels to resolve it [3, 64, 69].

Live and asynchronous formats therefore differ in when learners can seek clarification [3]. In live settings, learners can ask questions and receive feedback in the moment [3]. In contrast, when lectures are consumed asynchronously, learners lose direct engagement with instructors and peers, and questions or feedback are more likely to occur through delayed follow-up channels rather than during the viewing moment itself [3, 69]. As a result, asynchronous viewers often rely on replaying difficult segments, revisiting recordings selectively, or seeking clarification through separate resources and communication channels [36, 39, 64].

Prior work has explored several forms of support for asynchronous video learning. One line of work provides timeline-based navigation, annotation, and note-linked access to help learners revisit relevant parts of a video in context [23, 48, 53]. Another line of work aggregates collective viewer traces, such as comments, reactions, and playback behaviors, to make shared points of difficulty and discussion more visible over time [36, 40, 71, 74]. More recently, LLM-powered conversational systems have begun to provide on-demand explanation and interactive guidance during learning [7, 55, 66, 76].

However, these approaches still leave important gaps. Timeline-anchored comments and learning traces can make difficult moments easier to notice and revisit, but they often leave learners to interpret those traces on their own rather than turning them into direct support for the current moment [14, 40, 74]. Conversational agents can respond directly to learner questions, but existing work places more emphasis on question answering than on how explanations are grounded in presentation materials, local video context, or accumulated learner traces [7, 66, 76]. Taken together, prior work shows the value of making difficult moments visible and answerable, but less often connects trace-based visibility, context-grounded explanation, and instructor follow-up within a single support framework.

To address these, we present EchoMind, a system that augments recorded presentations with a context-grounded conversational agent and collective viewer signals. When viewers have a question, the video is automatically paused to allow them to ask questions of an agent that has access to the current position in the video, the presentation transcript, slides, supporting papers, and other materials the speaker provides, as well as dynamically accumulated viewer signals such as prior questions, reactions, and video behavioral traces. The system uses specialized retrieval-augmented generation (RAG) [41] to ground responses in the actual content and current viewing context. When the agent's confidence falls below a threshold, questions are asynchronously escalated to the speaker, with human-provided answers integrated back into the knowledge base for future viewers. EchoMind strengthens learner–content interaction through context-grounded explanation, supports learner–instructor interaction through human-in-the-loop escalation and speaker feedback, and enables an indirect, trace-mediated form of learner–learner interaction through accumulated viewer questions, reactions, and behavioral traces.

We conducted a between-subjects study with 75 participants recruited via Prolific [1], comparing three conditions: video-only, video with a generic LLM chatbot lacking access to presentation materials or collective viewer signals, and video with EchoMind (25 participants per condition). The participants watched a five-minute recorded lecture on computer science topics and interacted with their assigned system. EchoMind users reported significantly more positive learning experiences across multiple dimensions, suggesting that grounding agent responses in presentation materials and collective viewer signals meaningfully improves the subjective experience of asynchronous video learning in this study.

The contributions of this work are the following: (1) EchoMind, a system that augments recorded presentations with a context-grounded conversational agent, community awareness visualizations, human-in-the-loop escalation, and a speaker dashboard; (2) a RAG-based architecture that integrates presentation materials with viewer signals, enabling context-aware and presentation-aligned responses; (3) empirical findings from a 75-participant between-subjects study demonstrating that EchoMind significantly improved the subjective learning experience across multiple dimensions, while ungrounded LLMs failed to provide reliable benefits over video-only viewing; (4) a design rationale for reintroducing multiple forms of educational interaction into asynchronous video learning through accumulated questions, reactions, behavioral traces, and speaker clarifications (with the viewer-facing subset evaluated in this study).

## 2 Related Work

This section reviews prior work on asynchronous video learning, focusing on its key challenges and how prior systems leverage temporal interaction traces and LLM-based conversational approaches in attempts to support understanding.

### 2.1 Predicaments of Asynchronous Video Learning

With the growing adoption of online education and academic communication, recorded presentations have become a central medium in higher education, academic talks, and professional training. Recorded videos offer learners substantial temporal flexibility, allowing them to progress at their own pace and revisit complex concepts through pausing and replaying [36, 42]. Prior work documents distinct patterns of selective access and video interaction when learners use live and archived course materials [10, 26]. Despite these benefits, asynchronous viewing does not necessarily lead to smooth understanding [68]. More broadly, research on online learning at scale suggests that prerecorded and large-scale online learning environments often struggle to sustain learners' attention over time and can easily slide into passive consumption when interaction is limited [17]. Prior work suggests four recurring difficulties in asynchronous video learning: learners may get stuck at specific moments in the video, these moments can interrupt attention continuity, it is often hard to tell whether others struggled in similar ways, and immediate instructor clarification is usually unavailable.

*Local comprehension difficulties in video.* Research based on learning analytics and video interaction data consistently suggests that comprehension difficulties in video learning often surface around specific moments during viewing rather than as a uniform loss of engagement [26, 27, 68]. Prior work further indicates that playback behaviors such as pausing, rewinding, and seeking can provide useful signals about how learners are managing those moments [36, 39, 42, 72]. This is also consistent with research on instructional videos showing that learners must continuously process changing audiovisual information, which makes pauses and segmentation important for following and understanding the material [47].

*Attention continuity and illusion of understanding.* Further studies indicate that these breakdowns are local and temporally situated: learners may remain engaged overall while suddenly becoming confused at a particular concept, derivation, or example, or experience added cognitive load when navigating between a video and accompanying documents [19, 27, 36, 51]. When confusion is not adequately addressed, learners may develop an illusion of understanding, believing they have grasped the material despite persistent conceptual gaps [15]. In such moments, unresolved local confusion can also interrupt attention continuity, pushing learners away from active sensemaking toward passive viewing or disengagement [17].

*Limited social visibility.* Unlike face-to-face classrooms or live-streaming settings, learners cannot observe peers' struggles, pauses, or questions, making it difficult to contextualize their own moments of confusion [10, 20, 65]. Beyond education-specific systems, the social translucence framework argues that making socially significant activity visible can support awareness and accountability, enabling people to draw on social information to structure and coordinate their interactions in otherwise opaque digital spaces [21]. In video learning, time-anchored comments make peer reactions visible at relevant moments and can reshape attention allocation and perceived engagement during viewing [14, 40]. Without such

[1] https://www.prolific.com/

social references, learners struggle to distinguish common content difficulties from individual knowledge gaps, often falling back on repeated rewatching and other high-effort retrospective strategies [11, 27, 30, 42].

*Reduced opportunities for immediate instructor clarification.* Distance-education frameworks distinguish learner–content, learner–instructor, and learner–learner interactions as central interaction types [32, 49]. Recorded lectures primarily preserve the first of these, while immediate instructor-mediated clarification is substantially weakened. As a result, even when learners recognize a difficulty, they often have limited means to obtain timely explanation or feedback at the point where confusion occurs.

## 2.2 Temporal Learning Traces in Video-Based Learning

To address the difficulty of revisiting confusion that arises at particular moments in a video, prior work has proposed a range of time-aligned comments, annotations, and learning trace systems that bind questions, feedback, and notes to specific timestamps or segments [40, 48, 71]. One common form centers on comments and peer discussion tied to moments in the video, allowing later viewers to encounter prior learners' reactions when they reach the same point [40, 71]. Other systems attach notes, navigation aids, or prompts to specific parts of a video so that learners can revisit relevant content in context rather than through detached discussion alone [23, 48, 53]. Taken together, these systems re-anchor otherwise scattered learning activity to the video timeline, making difficult moments more recoverable and interpretable in place.

This class of systems contributes in at least three ways. First, it restructures discussion and feedback around the timeline, helping learners quickly locate relevant moments during review rather than search through disconnected forums or notes [40, 74]. Second, it increases social visibility in asynchronous learning by showing that others also commented, questioned, or lingered at similar points [40, 71]. Third, it can shape how learners allocate attention during viewing. Research on timeline-anchored comments shows that learners switch attention frequently between the video and the comments, and that layout choices affect switching frequency, retrieval speed, and learning outcomes [14]. Related systems also use time-synced traces to support collective review, restore a degree of social presence, or help instructors reflect on where learners struggle [22, 33, 44, 62, 74].

However, these systems also show clear limits. In time-anchored comment systems, later viewers still need to read prior comments and decide for themselves which ones are relevant to their current difficulty [40, 71]. CoKnowledge makes collective comments easier to browse by organizing them around the timeline, but its main contribution is still to support assimilation and review of collective knowledge rather than to generate direct help for the learner's current question [74]. EduLive similarly uses transcript-linked annotations to recover a sense of interaction in synchronous live learning [22], while TSConnect aggregates learner traces to help instructors reflect on where students struggle after class; in both cases, these traces primarily support awareness, reflection, and retrospective analysis, leaving immediate in-the-moment explanation less developed [44]. Existing work also suggests that time-aligned traces must be integrated carefully, since added comment layers can redirect attention during viewing, and different visual layouts can change switching frequency, retrieval speed, and learning outcomes [14]. Overall, these systems make difficult moments easier to notice, revisit, and discuss. Their limitation is that these traces are rarely connected to direct support grounded in the learner's current moment, the lecture materials, and the explanation being given.

## 2.3 Interactive Support in Video-Based Learning

If time-aligned comments and learning trace systems primarily help make difficult moments visible, conversational support systems take a further step by trying to answer questions about those moments when they arise. Although annotations and traces can help learners identify where to return, they do not usually answer the specific questions learners form while watching. Recent work has therefore introduced conversational agents and AI-based tutors into learning settings to support question asking [7], provide personalized guidance [66], and scaffold reflection through staged interaction [76].

The main contribution of these systems is that they turn relatively passive video viewing into a more interactive learning process. Conversational systems go beyond showing where others commented or which parts may matter by directly responding to learners' current questions with explanation or guidance. For example, SAM provides context-aware question answering around lecture content and positive user experiences and benefits for some learner groups compared with video-only baselines [7]. LearnMate offers personalized learning plans and guidance tailored to learner progress [66]. AutoPBL similarly uses LLM-based support to scaffold reflection and iterative sensemaking through staged interaction [76]. Taken together, these systems show that on-demand explanation can make learning more interactive and give learners more direct opportunities to ask for clarification during study [7, 66, 76].

However, several limitations remain. First, many conversational learning systems still depend on learners to recognize their own confusion and articulate it clearly before support can begin, so assistance remains largely reactive [7, 46, 58, 61, 76]. This leaves weaker support for moments of difficulty that remain implicit, fleeting, or only partially formed, such as when learners pause or rewind without being able to formulate a precise question [27, 36]. Second, the current literature emphasizes question asking and response generation, while offering less detail on how explanations are tied back to specific lecture materials, playback positions, or contextual evidence visible to the learner [7, 66, 76]. Third, these systems generally focus on providing automated support within the interface, while offering less discussion of how unanswered, ambiguous, or instructor-dependent questions should be handled [7, 66, 76].

## 2.4 Communication with Instructors and Human Handoff

Although time-aligned traces and conversational support can partially reduce comprehension difficulties in video learning, they do not eliminate the role of the instructor [22, 44, 57]. Learner–instructor interaction remains important for clarification, feedback, and continuity of support in online learning environments [1, 18,

32]. Work on online learning similarly shows that instructor communication and facilitation are closely tied to teaching presence and to learners' sense of support in online courses [1, 24, 43].

In asynchronous settings, this instructor support is often provided through email, discussion boards, forums, or office hours rather than through immediate question answering during viewing. Such channels allow learners to ask follow-up questions after the fact and give instructors a way to see which issues recur across learners [4, 18, 69]. Research on asynchronous online discussions further suggests that instructor participation, feedback practices, and discussion design can influence participation and performance, although the effects depend on factors such as class size, course goals, and the form of instructor involvement [18, 69]. Related review work on feedback in higher education also emphasizes that effective feedback depends on its integration with learners' ongoing activity as opposed to its availability as a separate resource [6]. These mechanisms therefore remain an important source of human support when automated or peer-based resources are insufficient.

However, existing instructor communication mechanisms still fit awkwardly with recorded presentation settings. In many online courses, learner–instructor communication is organized through separate channels such as discussion spaces, course communication structures, and feedback activities alongside the learning content itself [4, 6, 18]. These channels support follow-up questions and feedback, but asynchronous discussions necessarily separate some responses from the original viewing moment [69]. For recorded presentation settings, this suggests a practical gap: instructor support may remain available, yet it is not necessarily provided within the same moment of viewing in which the difficulty first arises.

This points to a further design problem: not every question should be handled by the system alone, yet existing work offers limited guidance on how automated support should connect to instructor involvement when AI responses are insufficient or when escalation is more appropriate. From this perspective, the challenge in asynchronous video learning is not only whether learners can contact an instructor, but also how support should be distributed between automated assistance and instructor intervention while still preserving moment-specific context for learners and aggregated insight for instructors.

Overall, prior work addresses asynchronous video learning from several complementary angles. Research on the predicaments of asynchronous learning identifies local comprehension difficulties, interruptions in attention continuity, limited social visibility, and reduced opportunities for immediate clarification. Time-aligned comments and learning trace systems improve the visibility and revisitation of difficult moments. Conversational support systems add on-demand explanation. Instructor communication and asynchronous discussion mechanisms provide a human support channel when system support is limited. Yet these forms of support often remain separate from one another: some make difficult moments visible, some answer questions, and some provide delayed instructor feedback. What remains underdeveloped is an integrated approach that can connect situated clarification, social visibility, material grounding, dual-role needs, and human fallback within the same support framework. The next section distills these unresolved tensions into EchoMind's design challenges.

## 3 Design Challenges: Conversational Agents for Online Videos

We derived several design challenges from prior literature and pilot observations (Figure 2). These challenges guided the design of our work and may inform future systems that augment video-based learning with conversational agents.

- **DC1 - Loss of Context.** Learners' confusion during educational video viewing is often immediate and tied to the specific context [27]. When they must pause the video, leave the interface to search for information, or reformulate a question, their attention is easily disrupted and the original learning flow is broken [29, 55]. This suggests that explanatory support should occur within the viewing context, rather than requiring users to switch interfaces or tasks.
- **DC2 - Lack of Community Awareness.** In traditional video-based learning, learners cannot see others' states of understanding and thus cannot easily judge whether their confusion is personal or common [57]. This lack of reference can make learners feel isolated, reduce self-efficacy, and increase unnecessary pressure when confronting complex content [43]. Both prior literature [8, 38] and our long-term observations of online learning show that lightweight behaviors, such as pausing, rewinding, adjusting playback speed, or using emoji reactions, although not formal questions, often reflect learners' momentary difficulty or shifts in attention.
- **DC3 - Lack of Reliability and Provenance.** Conversational agents in educational settings can produce responses that diverge from the speaker's intent, lack supporting evidence, or hallucinate content entirely [56, 75]. Without clear attribution to source materials, learners cannot verify whether an explanation accurately reflects the lecture content or stems from unconstrained model generation. This undermines trust and makes it difficult for learners to calibrate their confidence in system-provided answers [56].
- **DC4 - Competing Information Needs.** Speakers and learners have different priorities: learners need localized, in-the-moment understanding support [12], while speakers need aggregate patterns and questions requiring their attention [6]. Educational video learning also carries inherently high cognitive load, and excessive visual changes, frequent interface updates, or complex interactions may introduce additional distraction [22]. A single unified interface risks overwhelming both parties with information irrelevant to their immediate task.
- **DC5 - Limits of Automated Response.** Not all questions are suitable for model-only responses. Some involve speaker perspectives, clarifications of complex reasoning, pedagogical intentions, or suggestions related to presentation content [70]. Automated generation in such situations may produce inaccurate explanations, deviate from the speaker's intent, or fail to address learners' actual needs—yet systems rarely signal when they have reached the boundary of their competence.

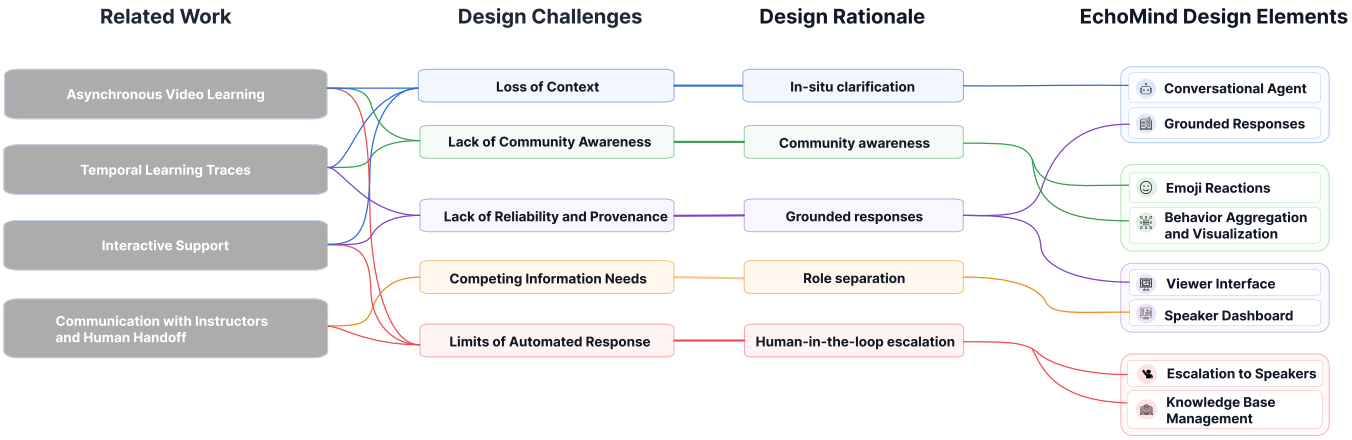



**Figure 2: Conceptual mapping from prior work to EchoMind design. Prior work in asynchronous video learning, temporal learning traces, interactive support, and instructor communication motivates EchoMind's design challenges. These challenges are translated into design rationales and operationalized as concrete design elements, including conversational agents, grounded responses, engagement visualization, viewer and speaker interfaces, escalation to speakers, and knowledge base management. Colored pathways indicate conceptual mappings between challenges, rationales, and system components.**

## 4 EchoMind

EchoMind is a web-based video player that augments recorded presentations with an LLM-powered conversational agent. Viewers watch a presentation and, at any moment, can pause to ask questions. The agent retrieves evidence from speaker-provided materials, the transcript, slides, and supplementary documents, using retrieval-augmented generation (RAG) [41]. When the agent cannot answer confidently, questions escalate to the original speaker, whose responses become part of the knowledge base for future viewers. The system also captures viewer behavior: pauses, rewinds, playback speed changes, and emoji reactions. These signals aggregate into visualizations that help both viewers (who see where others struggled) and speakers (who see which segments need clarification).

EchoMind separates functionality into two interfaces. The *viewer interface* () supports real-time engagement during playback: the conversational agent, emoji reactions, and visualizations of aggregate viewer behavior. The *speaker dashboard* () provides post-hoc analytics: engagement patterns, confusion hotspots, escalated questions awaiting response, and tools to augment the knowledge base. This separation ensures viewers focus on learning while speakers focus on improving content—neither is distracted by functionality meant for the other.

The remainder of this section describes how EchoMind addresses the design challenges from Section 3. We first explain how the system maps each challenge to a design response. We then describe the viewer interface and speaker dashboard in turn. Finally, we detail implementation choices. Figure 3 shows the system architecture.

### 4.1 Design Rationale

We address the design challenges (DC1–DC5) from Section 3 through five corresponding design responses:

**DC1 → In-situ clarification ()** Viewers ask questions directly within the playback interface. The agent's responses appear alongside the video, grounded in presentation content. This allows clarification without task-switching or leaving the viewing context.

**DC2 → Community awareness ()** Viewers express reactions using emoji during playback. The system aggregates these reactions across viewers and visualizes them on the timeline. This lets viewers see where others found content confusing or engaging, reducing the isolation of asynchronous learning.

**DC3 → Grounded responses ()** All agent responses cite specific sources: transcript segments, slide content, or supplementary materials. As questions and answers accumulate, they become part of the knowledge base, enabling consistent explanations and reflecting prior confusion patterns.

**DC4 → Role separation ()** Viewers and speakers have distinct interfaces tailored to their tasks. Viewers see the conversational agent, source citations, and aggregate reactions. Speakers see engagement analytics, confusion hotspots, and escalated questions. Neither interface contains distractions from the other role.

**DC5 → Human-in-the-loop escalation ()** When the agent detects insufficient evidence or low confidence, questions escalate to the speaker. Speakers review the agent's preliminary answer and supporting evidence, then revise or augment as needed. Approved responses integrate back into the knowledge base for future viewers.

### 4.2 Viewer Interface

The viewer interface combines a browser-based video player with four components: a conversational agent, an emoji reaction system, aggregate behavior visualizations, and an escalation mechanism.

*4.2.1 Conversational Agent.* The conversational agent (Figure 1-d) responds to viewer questions using a RAG-based mechanism (Figure 4). When a viewer focuses on the question input field, the video

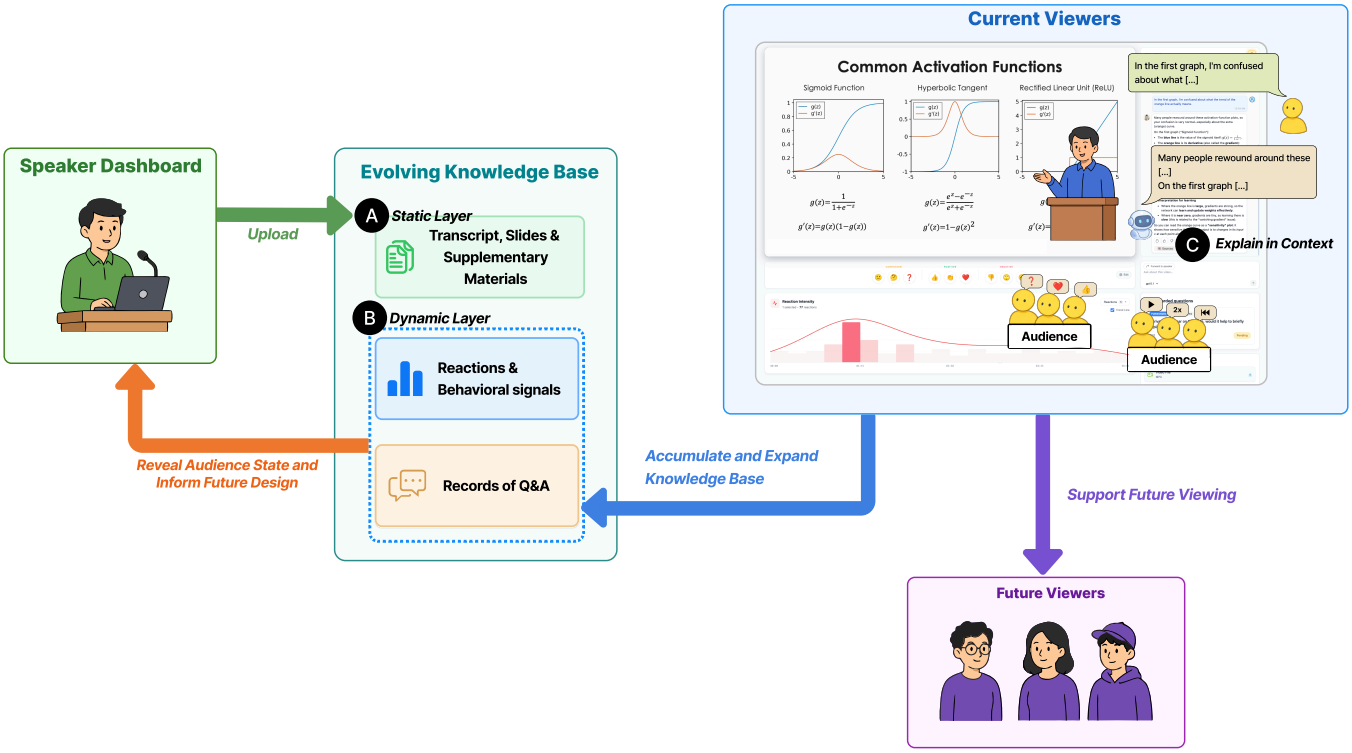



**Figure 3: EchoMind architecture. (a) Static layer. When a speaker uploads a lecture, the system constructs a static knowledge base consisting of the video transcript, slides, and supplementary materials. This layer remains stable after upload and serves as the primary retrieval source for question answering. (b) Dynamic layer. As viewers watch the video, their interaction traces—including playback behaviors (e.g., pauses, rewinds), reactions, and questions—are continuously logged and aggregated. In addition, resolved Q&A exchanges (including speaker-escalated answers) are stored in this layer, which grows over time. (c) In-context explanation during viewing. Viewers interact with a conversational agent embedded in the video player. The agent retrieves relevant content from the static layer and augments it with aggregated signals from the dynamic layer (e.g., where many viewers slow down or ask questions), providing explanations grounded in the current video context. Aggregated interaction traces and Q&A records are visualized along the video timeline for both speakers and future viewers, exposing collective patterns of difficulty and engagement that accumulate and shape subsequent viewing experiences.**

automatically pauses. The system then retrieves relevant content from the knowledge base, prioritizing segments near the current playback position, and captures the current video frame to incorporate visual context from slides (Figure 3-c). The agent generates a response constrained by the retrieved evidence and displays the supporting sources, allowing viewers to inspect the basis of each answer.

The agent classifies questions as content-related or off-topic using an implementation-specific prompt. It also assigns a confidence score, motivated by prior work on model confidence and confidence-based orchestration [25, 34]. Content-related questions with low-confidence answers trigger automatic escalation to the speaker.

To reduce the effort of formulating questions, the interface provides *recommended questions*—prompts derived from the knowledge structure around the current timestamp and common patterns of viewer uncertainty. These serve both as conversation starters and as lightweight self-assessment cues.

*4.2.2 Emoji Reactions.* The emoji reaction interface (Figure 1-a) provides a low-effort channel for viewers to report their cognitive state without interrupting playback. Drawing on prior emoji research [5], we use a design-specific taxonomy with three categories: *confusion* (cognitive breakdown), *positive* (understanding or agreement), and *negative* (disengagement or frustration). Each category has three default emoji:

- Confusion: 😕 🤔 ❓
- Positive: 👍 👏 ❤️
- Negative: 👎 🙄 😴

Viewers can remap these defaults to their preferred emoji. All reactions are timestamped and aggregated across viewers for visualization.

*4.2.3 Behavior Aggregation and Visualization.* Beyond explicit reactions, the system logs playback interactions as implicit behavioral signals: play/pause events, backward and forward seeks, and playback speed adjustments. We also treat question events as a distinct

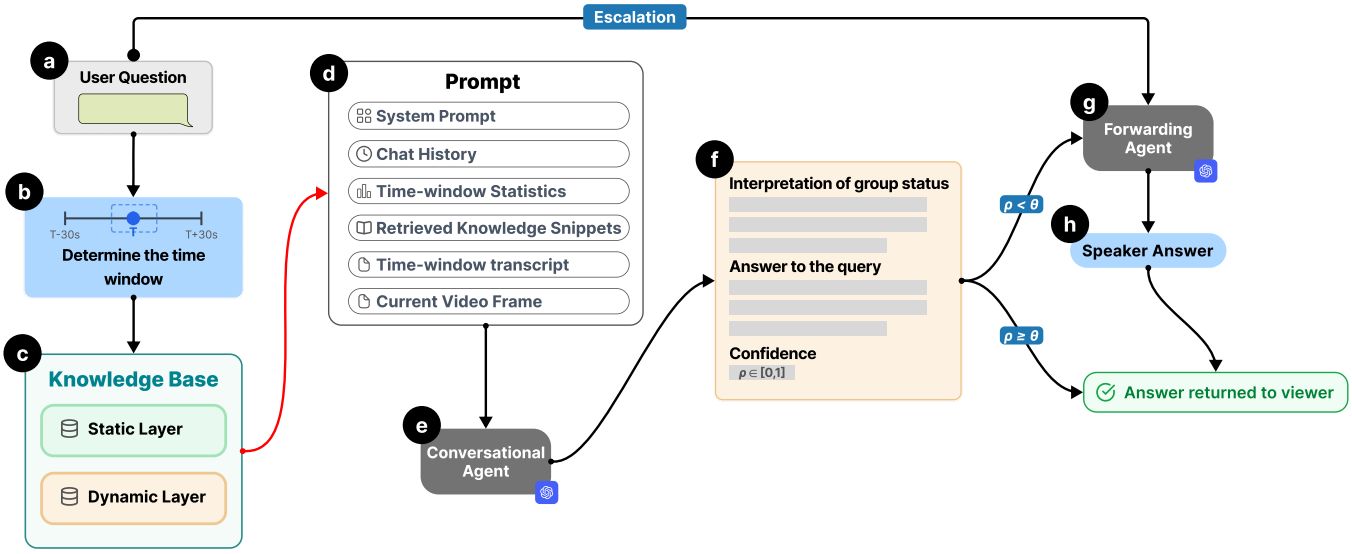



**Figure 4: RAG-based Q&A Pipeline. When a viewer submits a question (a), the system determines the relevant time window centered on the current playback timestamp (b) and retrieves context from the knowledge base (c) via two parallel paths. Time-window statistics, including emoji reactions and playback behavior data, are retrieved from the dynamic layer by timestamp matching. Transcript segments within a ±30-second window around the current timestamp provide local temporal context, while vector similarity search retrieves additional relevant transcript segments, slide content, and historical Q&A records from the static and dynamic layers. The retrieved context is assembled into a structured prompt (d), together with the system prompt, chat history, and current video frame, and passed to a conversational agent (e). The agent produces a structured response (f) that includes an interpretation of group engagement status, an answer to the query, and a confidence score ($\rho$). If the confidence score meets the threshold ($\rho \geq \theta$), the answer is returned directly to the viewer. Otherwise, it is routed to a forwarding agent (g) for quality filtering and escalation to the speaker (h), whose expert answer is subsequently returned to the viewer. Viewers may also manually trigger escalation at any time (top).**

form of pause. The behavioral signals, combined with emoji reactions and question events, aggregate into two visualizations.

The **Reaction Intensity View** (see Figure 1-b) displays the distribution of emoji reactions along the timeline as a bar chart aligned with the playback axis. The system partitions the video into adaptive temporal bins (finer for shorter videos, coarser for longer ones) and accumulates reactions within each bin. An optional kernel density estimation curve overlays the bars to show global trends. Viewers can click any bar to jump to that moment in the video.

The **Cognitive Curve** (see Figure 1-g) integrates all behavioral signals, including pauses, seeks, speed changes, emoji reactions, and questions, into a single confusion-oriented trend line. Inspired by prior work that quantifies user behavior through weighted interaction signals [59], we conducted a formative study ($N = 16$) to derive signal weights using direct relevance ratings. Participants rated the extent to which each signal reflected confusion. To account for individual differences in scale usage, ratings were normalized within participant and then aggregated across participants. The resulting values were used as empirically derived signal weights for computing the Cognitive Curve (see Table 1). The curve appears near the video player and collapses when playback controls are hidden to minimize distraction.

At peak regions of the cognitive curve, the system generates **Context Notes**: brief summaries identifying potential sources of confusion. Notes appear in a collapsible sidebar and surface only those near the current playback position, serving as lightweight guidance rather than additional reading burden.

**Table 1: Behavioral signal weights. Operationalized behavior weights derived from our 16-participant formative study rating each signal's relevance to confusion. Behaviors are sorted in descending weight.**

| Signal | Weight |
|---|---|
| Backward seek (rewatch) | 0.190 |
| Pause | 0.172 |
| Confusion emoji | 0.153 |
| Slow down | 0.150 |
| Negative emoji | 0.121 |
| Speed up | 0.078 |
| Positive emoji | 0.070 |
| Forward seek | 0.067 |

*4.2.4 Escalation to Speakers.* Questions escalate to speakers in two ways: automatically (when the agent's confidence falls below threshold) or manually (when viewers explicitly request human

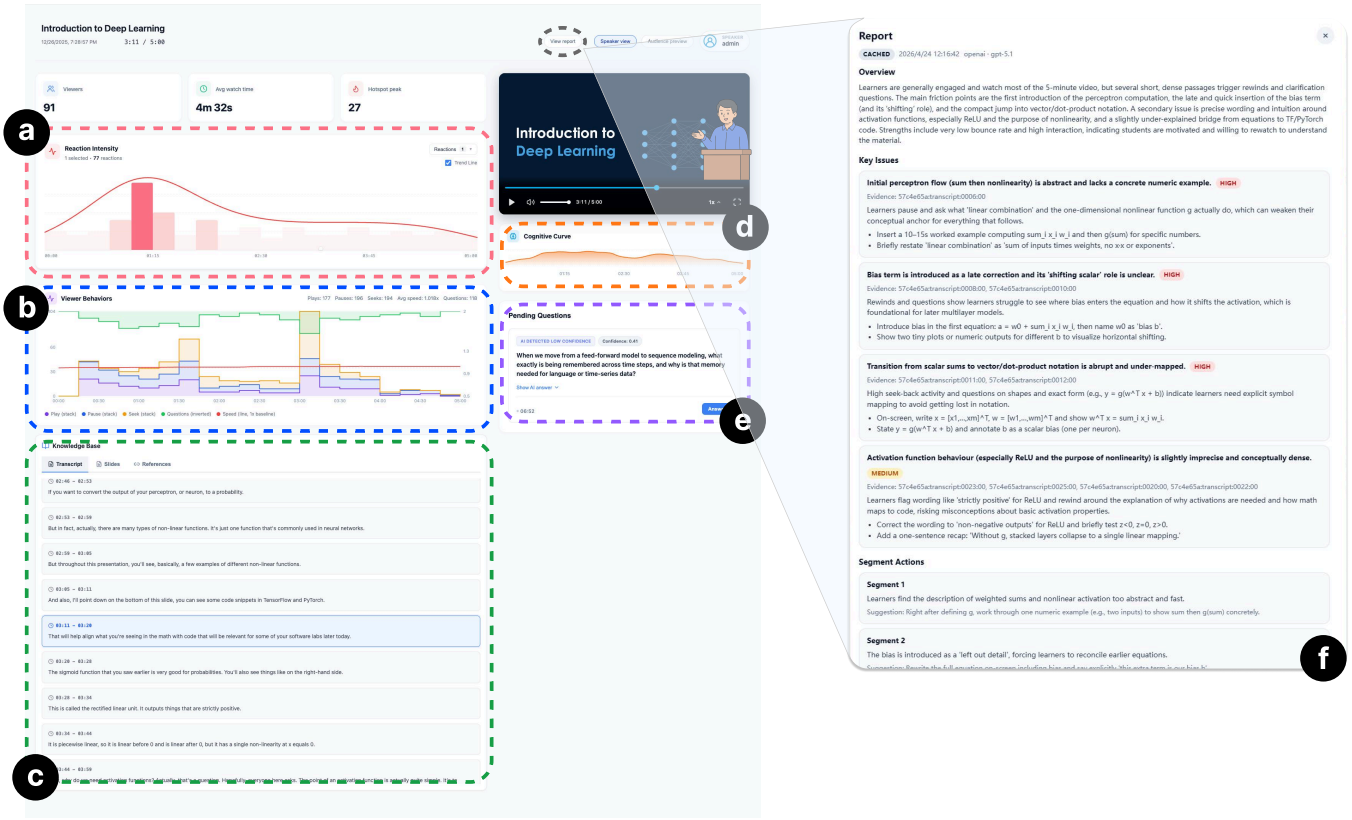



**Figure 5: Speaker-facing dashboard in EchoMind. (a) Reaction intensity visualizes collective learner reactions over time. (b) Viewer behaviors summarize aggregated playback actions such as pauses, rewinds, and revisits. (c) The knowledge base organizes lecture materials, generated explanations, and accumulated questions as grounded support resources. (d) The cognitive curve presents changes in estimated engagement or difficulty across the video. (e) Pending questions surface unresolved issues that require speaker attention. (f) The report summarizes engagement patterns, difficult moments, common questions, and suggested follow-up actions.**

clarification). Manual escalation is available via a toggle on the input field or an icon beneath the agent's response (see Figure 1-f).

To balance speaker workload with viewer expectations, a filtering agent determines whether a question is suitable for escalation. Questions deemed appropriate are forwarded to the speaker dashboard; others receive explanatory feedback suggesting alternative approaches. The escalation package includes the agent's preliminary answer, allowing speakers to revise rather than start from scratch.

## 4.3 Speaker Dashboard

The speaker dashboard provides post-hoc analytics linking engagement evidence to actionable improvements. The main area presents analytics and the knowledge base; a sidebar keeps the video player accessible for precise navigation.

*4.3.1 Engagement Analytics.* Speakers see aggregate engagement metrics: total viewers, average watch time, and completion rates. The dashboard surfaces several visualizations:

**Reaction Intensity (see Figure 5-a)** Distribution of confusion, positive, and negative reactions along the timeline, with optional trend overlay.

**Cognitive Curve (see Figure 5-d)** The same confusion-oriented trend line viewers see, revealing comprehension fluctuations across the audience.

**Viewer Behaviors (see Figure 5-b)** Integrated traces of playback behavior: play counts, pause counts, seek activity, and playback speed.

All visualizations are time-linked: clicking any point jumps to that moment in the video.

*4.3.2 Knowledge Base Management.* A knowledge recap panel displays the transcript, slides, and supplementary materials aligned to the current timestamp (see Figure 5-c). This allows speakers to cross-reference audience signals with the content that triggered them.

The escalation queue (see Figure 5-e) consolidates questions awaiting speaker response, both automatically escalated low-confidence queries and viewer-forwarded requests. For each question, speakers see the agent's preliminary answer and supporting

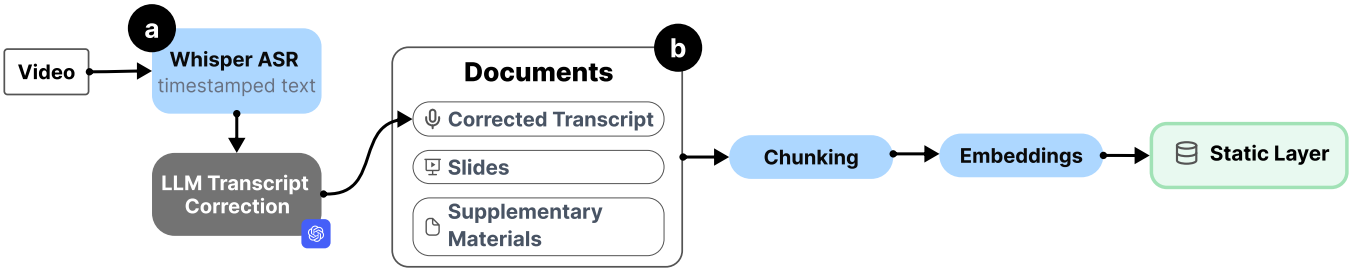



**Figure 6: Pipeline for Static Layer Construction.**

evidence. Approved responses integrate into the knowledge base for future viewers.

*4.3.3 Report Generation.* Speakers can generate a consolidated report that synthesizes collective signals (confusion hotspots, representative questions, playback patterns) into actionable feedback (Figure 5-f). The report identifies recurring issues and suggests concrete remedies.

## 4.4 Knowledge Base Construction

EchoMind maintains a two-layer knowledge base that evolves over time, consisting of a static foundation constructed at upload time (Figure 3-a) and a dynamic layer (Figure 3-b) that grows through viewer interaction.

*4.4.1 Static Layer Construction.* When a speaker uploads a video, the system constructs a knowledge base through automatic processing (Figure 6). The video's audio is transcribed to produce timestamped text (Figure 6-a). An LLM-based refinement module then corrects recognition errors, merges segments into semantic units, and aligns timestamps precisely. In parallel, the system parses speaker-provided slides and supplementary materials (e.g., related papers, extended readings). All content (Figure 6-b) is chunked and encoded into dense vector embeddings, which are stored in the static knowledge base for subsequent retrieval during viewer interactions.

*4.4.2 Dynamic Knowledge Base Expansion.* As viewers engage with the video, the knowledge base grows (Figure 7). Questions, agent responses, escalated answers from speakers, and aggregate reaction patterns all become retrievable context for future queries (Figure 3-b). This aims to create a form of emergent collaboration [63]: early viewers' questions benefit later viewers who encounter the same confusion.

## 4.5 Implementation

EchoMind consists of a backend server managing data and LLM pipelines, and a frontend web interface. For the frontend, the viewer interface and speaker dashboard are implemented in JavaScript using React. Because emoji appearance may vary across operating systems and fonts, and because users' interpretations of emoji can diverge [5], we adopt a unified visual style using the Noto Color Emoji font [2]. For the backend, the server is implemented in Python using FastAPI, with data stored in PostgreSQL. REST APIs provide core functionality for video management, question handling, and analytics aggregation.

The system integrates OpenAI's Responses API [3] for LLM inference. We use GPT-5.1 for conversational responses to viewers' questions. For tasks requiring lower latency and reduced token usage—transcript refinement, context note generation, and question classification—we use GPT-5-nano. Automatic speech recognition uses OpenAI's Whisper model.

# 5 User Study

We conducted a crowdsourced study to evaluate the core functions in EchoMind. Prior to the main study, we ran a pilot study (N=6) to refine the study procedure and seed the system with initial interaction traces. From the pilot study, we found that context notes would introduce additional experimental complexity beyond the scope of this study and that escalation to speakers requires real involvement from the original presenter; we therefore disabled these two components. The dynamic layer was initialized with behavioral data and reactions collected from pilot participants as accumulated viewer signals. To ensure all participants within each condition experienced identical system states, we disabled dynamic knowledge base expansion during the study; all participants interacted with the same knowledge base initialized from pilot study traces. All other viewer-facing mechanisms remained active. Accordingly, our study focuses on how EchoMind supports individual learners during recorded-presentation viewing.

> **RQ:** How does EchoMind reshape learners' experience of navigating the learning process during asynchronous video learning?

To operationalize the question, we formulate the following hypotheses:

- **HP1.** Compared to baselines, EchoMind provides greater support for addressing the in-situ needs of learners.
- **HP2.** Compared to baselines, EchoMind fosters a more active and supported learning experience.
- **HP3.** Ungrounded conversational AI fails to reliably improve difficulty resolution or the learning process compared to video-only viewing.

## 5.1 Experimental Conditions

Unlike prior evaluations of video-based learning assistants (e.g., Bodonhelyi et al. [7]), we include a condition with a generic LLM

[2] https://fonts.google.com/noto/specimen/Noto+Color+Emoji

[3] https://developers.openai.com/api/docs

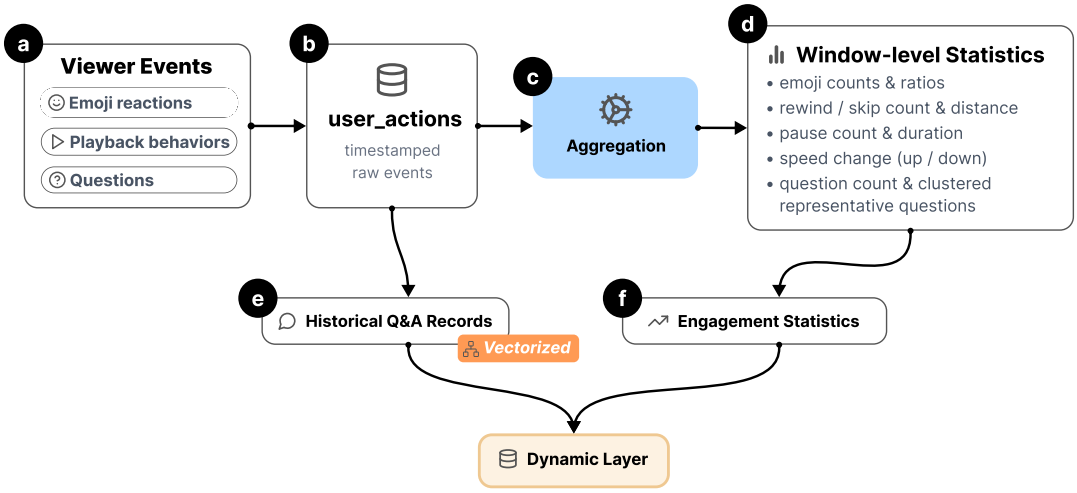



**Figure 7: Pipeline for Dynamic Knowledge Base Expansion. As viewers watch the video, (a) three types of behavioral events, including emoji reactions, playback interactions (pause, rewind, forward skip, and speed change), and natural language questions, are continuously logged as timestamped records in the (b) `user_actions` table. These raw events are periodically processed by an (c) aggregation module that segments the video timeline into semantically defined time windows based on the refined transcript and computes (d) per-window statistics, including emoji counts and normalized ratios across three sentiment categories (confusion, positive, and negative), rewind and skip counts and distances, pause count and duration, speed change counts, and clustered representative questions. The resulting (f) engagement statistics are indexed by time window for direct timestamp-based retrieval. In parallel, resolved (e) Q&A records, including expert answers provided by the speaker, are encoded as vector embeddings to support semantic retrieval. Both components are stored in the dynamic knowledge base, which grows continuously as more viewers interact with the video.**

agent lacking access to presentation materials, isolating the effect of material grounding. Accordingly, we compared three conditions in a between-subjects design:

**VideoOnly (G1)** Standard video playback with no interactive support. This represents the baseline online learning experience.

**QAOnly (G2)** Video playback with a generic LLM-powered chatbot. The agent can answer questions but has no access to the video transcript, slides, or supplementary materials.

**EchoMind (G3)** Video playback with the full EchoMind conversational agent. The agent uses RAG to ground responses in the video transcript, slides, and related materials, and displays source provenance for each answer.

All participants watched the same 5-minute video segment: the introduction to perceptrons from a recorded lecture titled *Introduction to Deep Learning* [4].

All three conditions were implemented on the same interface: they shared an identical video player, layout, and visual styling, and differed only in which components were visible. VideoOnly hid all interactive components; QAOnly hid the emoji reaction bar, visualizations, and escalation controls while retaining the chat panel, which used the same front-end component as EchoMind and differed only in the underlying agent.

[4]https://youtu.be/alfdI7S6wCY?si=_eCOk9mM9P_UUw9r&t=1161

## 5.2 Procedure

Figure 8 shows the study flow. Participants logged into the study system and were randomly assigned to one of the three conditions. They first completed a pre-task questionnaire assessing demographic information and prior knowledge of deep learning. A brief tutorial then familiarized them with their assigned interface.

During the main session, participants watched the video and interacted with their assigned system. To encourage authentic engagement with the interactive features, participants in the QAOnly and EchoMind conditions were guided to ask at least one question during viewing, and EchoMind participants were additionally guided to record at least one emoji reaction. VideoOnly participants simply watched and completed a quiz. After viewing, all participants completed a comprehension quiz and post-task questionnaire.

We instructed participants not to use external tools (search engines, AI chatbots, Wikipedia) during the study. We could not enforce this restriction technically, but compliance was incentivized through task framing and time constraints.

## 5.3 Participants

We recruited 75 participants through Prolific [54], with 25 per condition. Ages ranged from 24 to 40 years ($M$ = 32.4, $SD$ = 4.38, $Mdn$ = 32). The sample included 38 male and 37 female participants, all reporting English as their primary language. Regarding educational attainment, 51 participants (68.0%) held undergraduate degrees, 18 (24.0%) held master's degrees, and 6 (8.0%) held doctoral degrees. The participants mainly resided in the United States

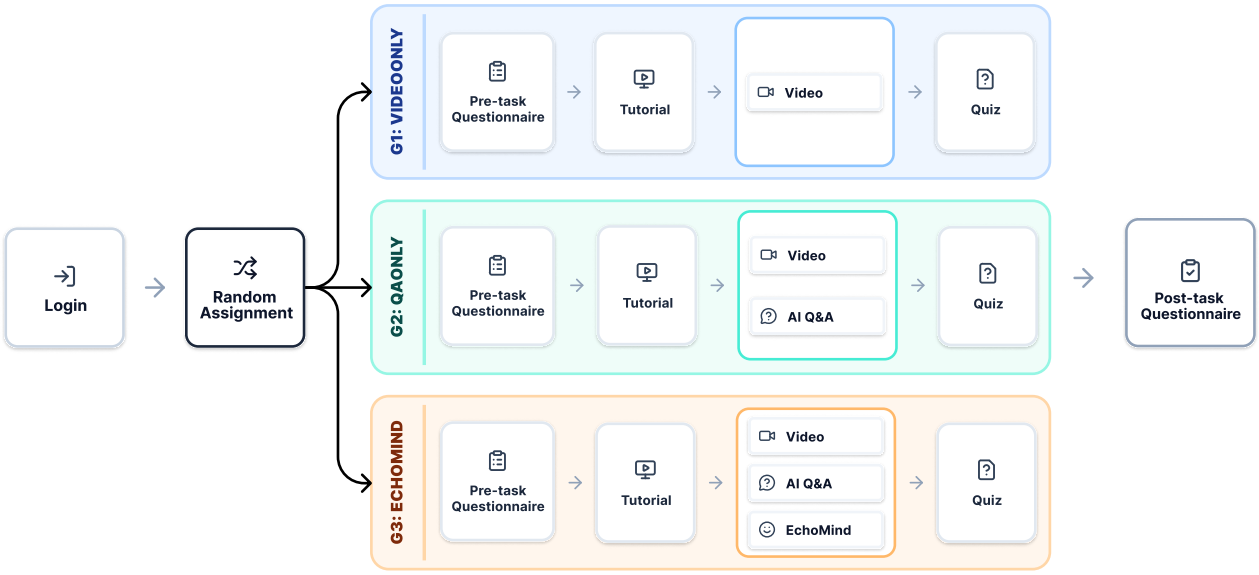



**Figure 8: Study procedure. Participants were randomly assigned to one of three conditions (VideoOnly, QAOnly, or EchoMind) and completed a pre-task questionnaire, tutorial, video-based learning session with condition-specific support, comprehension quiz, and post-task questionnaire.**

($n$ = 37, 49.3%) and the United Kingdom ($n$ = 36, 48.0%), with a small number in Ireland ($n$ = 2, 2.7%). Mean completion times recorded by Prolific were 14.04 minutes ($SD$ = 6.53) for VideoOnly, 15.27 minutes ($SD$ = 8.16) for QAOnly, and 16.93 minutes ($SD$ = 8.75) for EchoMind. Compensation was €12.79 per hour.

## 5.4 Measurement

We assessed learning outcomes through a comprehension quiz administered after viewing. We assessed subjective experience through a post-task questionnaire using 7-point Likert scales (1 = Strongly disagree / Not at all; 7 = Strongly agree / Extremely). The questionnaire covered four dimensions:

*Usability.* For usability of the system, referring to [9, 44], we designed these items:

- Ease of use;
- Learning support;
- System satisfaction; and
- Likelihood of future use.

*Workload.* Adapted from NASA-TLX [31], these items were assessed:

- Perceived mental demand;
- Perceived success in completing the task;
- Effort required to perform well; and
- Feelings of discouragement, stress, or frustration.

*Learning experience.* To examine how EchoMind supported participants' learning experience during video viewing, we assessed the following aspects, informed by evaluation approaches used in TSConnect and Glancee [44, 45] and closely related work:

- Clarifying confusing points [44, 45];
- Sustaining engagement during learning [37, 45];
- Encouraging reflection and active thinking [13, 44];
- Supporting moment-to-moment expression [44, 52];
- Fostering a sense of learning with others [33, 65]; and
- Promoting active rather than passive learning [7, 16].

*Feature-specific perceptions (EchoMind only).* For participants with access to EchoMind, referring to the feature-specific usefulness batteries used in prior system evaluations [44, 45], we designed additional items to assess perceptions of:

- Emoji reactions supporting reflective understanding;
- Cognitive curve revealing struggle points;
- Reaction chart conveying overall learning patterns;
- Usefulness of AI responses grounded in video materials;
- Usefulness of AI responses conditioned on user behavior; and
- Emoji reactions facilitating active participation.

# 6 Results

We report results across four dimensions: usability, workload, learning experience, and system-specific perceptions. Overall, EchoMind outperformed both baseline conditions on learning experience measures, with significant improvements in confusion clarification, expression support and sense of co-learning. Usability was comparable across conditions, though EchoMind showed higher intended future use. Workload did not differ significantly, though QAOnly exhibited the highest mental demand. EchoMind-specific features (cognitive curve, reaction visualizations, grounded AI responses) were rated positively by participants. Figure 9 summarizes the results.

We first assessed normality using the Shapiro–Wilk test; ratings were not normally distributed. We therefore employed nonparametric Kruskal–Wallis tests for omnibus comparisons. When

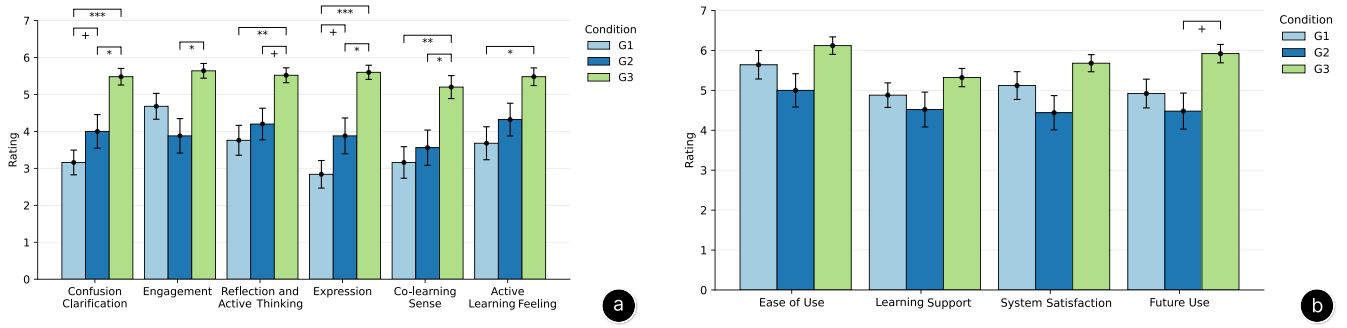



**Figure 9: Subjective ratings across conditions. (a) Learning process-related measures, including confusion clarification, engagement, reflection, expression, co-learning, and active learning. (b) Usability-related measures, including ease of use, satisfaction, perceived learning support, and future use. Significance levels are marked above bars. (+: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < .001$)**

Kruskal–Wallis indicated a significant difference across conditions, we used Dunn's post-hoc test with Holm–Bonferroni correction for pairwise comparisons.

## 6.1 Usability

We compared the usability of the three experimental conditions using post-task questionnaire data. Figure 9 shows participants' average ratings on usability-related items. A Kruskal–Wallis test revealed a significant group difference in *intended future use* ($H(2) = 6.00, p < .05$), whereas no significant differences were found for *ease of use* ($H(2) = 3.75, p = .153$), *support for the learning process* ($H(2) = 1.34, p = .511$), or *overall satisfaction* ($H(2) = 3.59, p = .166$).

Post-hoc Dunn tests further clarified the pattern for intended future use. EchoMind ($M = 5.92, SD = 1.15$) received higher ratings than QAOnly ($M = 4.48, SD = 2.26$), with this difference approaching significance after Holm correction (adj. $p = .068$). A similar though non-significant trend was observed between EchoMind and VideoOnly ($M = 4.92, SD = 1.80$; adj. $p = .110$). No difference was found between QAOnly and VideoOnly (adj. $p = .717$).

Although the remaining usability dimensions did not reach statistical significance, the mean ratings revealed a consistent pattern across measures. QAOnly received lower scores on *ease of use* ($M = 5.00, SD = 2.08$), *overall satisfaction* ($M = 4.44, SD = 2.14$), and *perceived learning support* ($M = 4.52, SD = 2.18$) compared with both EchoMind (ease of use: $M = 6.12, SD = 1.09$; satisfaction: $M = 5.68, SD = 1.07$; learning support: $M = 5.32, SD = 1.14$) and VideoOnly (ease of use: $M = 5.64, SD = 1.78$; satisfaction: $M = 5.12, SD = 1.74$; learning support: $M = 4.88, SD = 1.54$). In addition, QAOnly exhibited larger standard deviations across these measures, indicating greater variability in user experience.

## 6.2 Workload

Across the four workload measures, none of the Kruskal–Wallis tests reached conventional levels of statistical significance, including *perceived success* ($H(2) = 4.22, p = .121$), *effort* ($H(2) = 2.17, p = .339$), and *frustration and stress* ($H(2) = 0.21, p = .902$). *Mental demand* showed a near-significant trend ($H(2) = 4.71, p = .095$).

Although none of the differences reached statistical significance, descriptive trends were observed. For *mental demand*, QAOnly ($M = 6.20, SD = 0.91$) reported the highest workload, followed by VideoOnly ($M = 5.80, SD = 1.44$), with EchoMind showing the lowest values ($M = 5.32, SD = 1.52$). EchoMind ($M = 3.96, SD = 1.51$) received the highest *perceived success* ratings, compared to QAOnly ($M = 3.24, SD = 1.90$) and VideoOnly ($M = 3.12, SD = 1.64$). For *effort*, QAOnly ($M = 6.28, SD = 1.10$) showed higher scores than VideoOnly ($M = 6.08, SD = 1.04$) and EchoMind ($M = 5.88, SD = 1.13$). In contrast, *frustration and stress* remained comparable across the three conditions (EchoMind: $M = 4.36, SD = 1.68$; QAOnly: $M = 4.36, SD = 2.08$; VideoOnly: $M = 4.52, SD = 1.94$).

Overall, EchoMind exhibited workload levels comparable to or more favorable than VideoOnly across all dimensions, whereas QAOnly showed the highest mean levels on both mental demand and effort.

## 6.3 Learning Experience and Quiz Outcomes

*6.3.1 Clarifying points of confusion.* This item reflects the extent to which the system supported learners in clarifying moments of misunderstanding during the video ("The system helped me clarify points I was confused about").

A Kruskal–Wallis test showed a significant group difference ($H(2) = 17.17, p < .001$). Post-hoc Dunn tests indicated that EchoMind ($M = 5.48, SD = 1.12$) was rated higher than QAOnly ($M = 4.00, SD = 2.27$; adj. $p < .05$) and also higher than VideoOnly ($M = 3.16, SD = 1.68$; adj. $p < .001$), whereas QAOnly and VideoOnly did not differ (adj. $p = .099$).

*6.3.2 Expressing questions and emotions.* This item measures whether learners felt supported in expressing questions or emotions at the moment they occurred ("The system encouraged me to express questions or emotions at the moment they occurred").

A significant group difference was observed ($H(2) = 19.69, p < .001$). Post-hoc Dunn tests showed that EchoMind ($M = 5.60, SD = 0.96$) was rated higher than QAOnly ($M = 3.88, SD = 2.42$; adj. $p < .05$) and higher than VideoOnly ($M = 2.84, SD = 1.86$; adj. $p < .001$), whereas QAOnly and VideoOnly did not differ.

*6.3.3 Sense of learning together.* This item reflects whether participants felt they were learning together with others rather than alone ("I felt like I was learning together with others, not just by myself").

A significant group difference was found ($H(2) = 11.16, p < .01$). Post-hoc Dunn tests indicated that EchoMind ($M = 5.20, SD = 1.55$) scored higher than QAOnly ($M = 3.56, SD = 2.38$; adj. $p < .05$) and VideoOnly ($M = 3.16, SD = 2.13$; adj. $p < .01$), while QAOnly and VideoOnly did not differ (adj. $p = .514$).

*6.3.4 Reflecting and thinking actively.* This item assesses whether the system encouraged reflective and active thinking during learning ("The system encouraged me to reflect and think actively during learning").

A significant group difference was observed ($H(2) = 9.50, p < .01$). Post-hoc Dunn tests showed that EchoMind ($M = 5.52, SD = 1.00$) was rated higher than VideoOnly ($M = 3.76, SD = 2.03$; adj. $p < .01$), whereas the difference between EchoMind and QAOnly ($M = 4.20, SD = 2.14$) did not reach significance (adj. $p = .074$). QAOnly and VideoOnly also did not differ.

*6.3.5 Staying engaged during the video.* This item captures sustained engagement during viewing ("The system helped me stay engaged during the video").

A significant group difference was found ($H(2) = 7.51, p < .05$). Post-hoc Dunn tests indicated that EchoMind ($M = 5.64, SD = 0.99$) was rated higher than QAOnly ($M = 3.88, SD = 2.33$; adj. $p < .05$), whereas no significant differences were observed between EchoMind and VideoOnly ($M = 4.68, SD = 1.75$) or between QAOnly and VideoOnly.

*6.3.6 Overall shift toward active learning.* This item reflects whether the overall learning process felt more active rather than passive ("Overall, my learning process felt more active rather than passive").

A significant group difference was observed ($H(2) = 8.22, p < .05$). Post-hoc Dunn tests showed that EchoMind ($M = 5.48, SD = 1.19$) received higher ratings than VideoOnly ($M = 3.68, SD = 2.23$; adj. $p < .05$), whereas the difference between EchoMind and QAOnly ($M = 4.32, SD = 2.21$) was not significant. QAOnly and VideoOnly also did not differ.

*6.3.7 Quiz Outcomes.* Quiz performance did not differ significantly across conditions ($H(2) = 0.06, p = .972$). Descriptively, EchoMind achieved the highest mean score ($M = 3.04, SD = 2.89$), followed by QAOnly ($M = 2.96, SD = 2.46$) and VideoOnly ($M = 2.64, SD = 1.98$). The absence of significant differences is consistent with the study task's short duration and limited ecological validity, and does not preclude the possibility of comprehension gains in longer-term deployment contexts.

## 6.4 Perceptions of System

To describe participants' perceptions of EchoMind's system-specific mechanisms, we conducted one-sample Wilcoxon signed-rank tests against the neutral midpoint of 4 for all EchoMind-specific post-test items in the EchoMind condition ($n = 25$). All item means were above the neutral point, indicating generally positive perceptions. Descriptive statistics of the ratings are presented in Figure 10.

For *expressive reflection support* ("Seeing others' reactions and using emoji reactions helped me reflect on my understanding rather than passively watching"), ratings were above the neutral value ($M = 4.64$, $Median = 5.0$), but this deviation only showed a trend ($W = 67.0, p = .085$).

The remaining five mechanisms were all rated significantly above the neutral midpoint. Ratings for the *difficulty salience curve* ("The Cognitive Curve helped me notice where learners tended to struggle") were high ($M = 5.48$, $Median = 5.0$) and significantly different from 4 ($W = 12.0, p < .001$). Similarly, the *reaction chart for collective learning state* ("The Reaction Intensity chart helped me understand the overall learning status of the audience") showed a high median ($M = 5.56$, $Median = 6.0$) and a significant deviation from the neutral point ($W = 17.0, p < .001$).

For AI-related mechanisms, both *video-grounded AI responses* ("AI responses grounded in the video materials helped me understand the content better"; $M = 5.20$, $Median = 5.0$) and *signal-aware AI responses* ("AI responses that considered user behavior were useful"; $M = 5.40$, $Median = 5.0$) were rated significantly above the neutral midpoint ($W = 10.0, p < .001$; $W = 0.0, p < .001$, respectively).

Finally, *emoji-based participatory engagement* ("Using emoji reactions made me feel like I was actively participating, not just watching the video") was also rated significantly above the neutral point ($M = 5.48$, $Median = 6.0$; $W = 33.0, p < .01$).

Overall, with the exception of expressive reflection showing only a trend, all other EchoMind-specific mechanisms were perceived significantly more positively than a neutral evaluation.

# 7 Discussion

Our results lead us to accept HP1, partially support HP2 and HP3, revealing that the perceived value of AI in video learning depends critically on its contextual integration. Validating HP1, EchoMind substantially raised ratings for supporting in-situ expression compared to VideoOnly ($M = 5.60$ vs. 2.84). We note, however, that this difference partly reflects the availability of the emoji reaction channel in the EchoMind condition, and should not be attributed to contextual grounding alone. We partially support HP2: while EchoMind fostered a distinctive sense of social support ($M = 5.20$), the shift toward active thinking did not statistically exceed QAOnly. Workload did not differ significantly across conditions, but the descriptive pattern, with the highest mental demand under QAOnly and the lowest under EchoMind, points to a qualitative difference in how the two forms of AI support were experienced, which we examine in Section 7.4. Finally, we partially support HP3. Consistent with the hypothesis, the ungrounded agent failed to yield statistically significant improvements in confusion clarification or engagement compared to VideoOnly baseline. While it also elicited the highest mental demand and numerically lower future use intentions ($M = 4.48$ vs. 4.92), these negative impacts were observed as trends rather than statistically significant detriments. This suggests that without context, conversational AI may function as cognitive noise, imposing friction without delivering reliable experiential benefits.

## 7.1 Reframing Asynchronous Video Learning

EchoMind participants reported more positive experiences on several process-related measures, particularly confusion clarification,

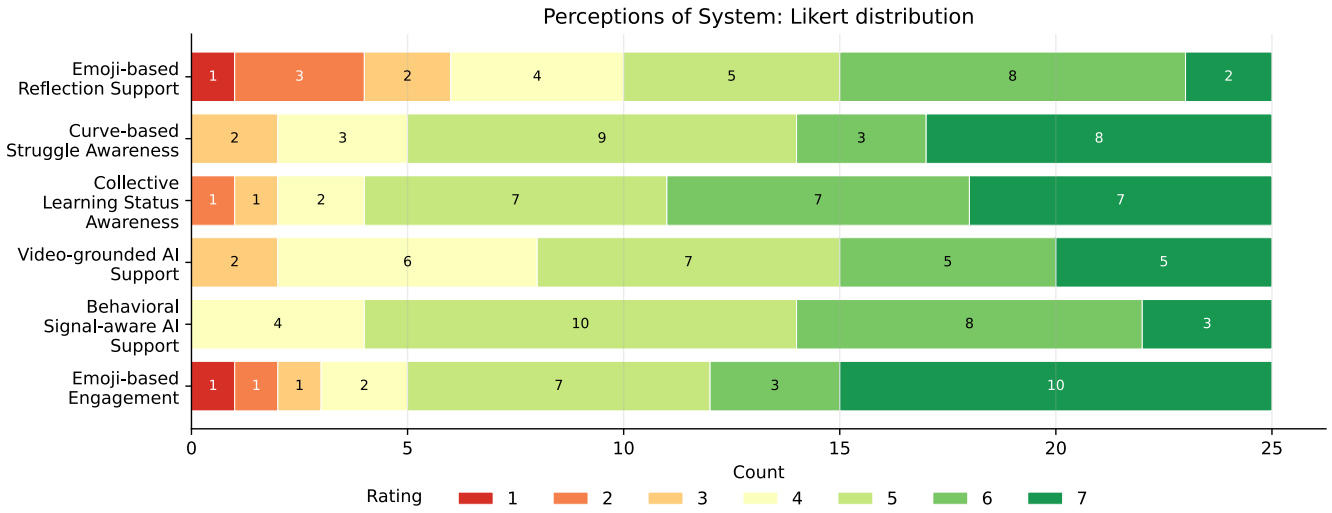



**Figure 10: Likert-scale distributions for system-specific perceptions in the EchoMind condition. Stacked bars show the distribution of participant ratings (1–7) across six items.**

in-the-moment expression, and co-learning. Prior work has shown that learners frequently experience localized breakdowns during video learning, which are typically addressed through rewatching or external search [30, 36, 42]. These strategies provide flexibility but separate clarification from the original viewing moment and from synchronous peer interaction [3, 10, 69]. EchoMind reframes these breakdowns as shared learning moments that can be accumulated and revisited over time, transforming individual confusion into collective learning resources.

More specifically, EchoMind significantly outperformed both QAOnly and VideoOnly in perceived support of clarifying confusion and supporting in-the-moment expression, while the latter two did not differ reliably. This finding further suggests that conversational AI is most useful when it is contextually integrated into video platforms [7, 14, 36, 39].

Beyond individual clarification, EchoMind also reshapes the social dimension of asynchronous learning. EchoMind aligns collective reactions and behavioral traces directly with the video timeline. This alignment allows learners to situate their understanding within a broader social context, consistent with prior work on social navigation and time-anchored peer interaction [21, 40, 71].

These results suggest that asynchronous video learning can be meaningfully reframed as a socially grounded process in which even static displays of prior learners' traces shape in-session interpretation, a reframing we view as a starting point rather than a conclusion about cumulative effects.

## 7.2 Why Contextualized AI Matters

Our results show that conversational AI becomes valuable in video-based learning when it is contextually integrated. Although QAOnly introduced an LLM-powered question-answering interface, it did not reliably outperform VideoOnly in supporting confusion clarification or in-the-moment expression. This finding challenges the increasingly common assumption that conversational AI alone constitutes meaningful learning support. Prior work on AI tutors and learning assistants has reported positive effects on engagement and perceived usefulness [7, 28, 56]. However, many such systems require learners to reconstruct context or formulate decontextualized questions, placing additional cognitive demands on learners. Our results suggest that without access to learning materials and temporal context, AI responses are prone to vagueness or misalignment, which may reduce their perceived usefulness and learners' willingness to continue using the system (consistent with QAOnly's lowest future-use ratings).

By contrast, EchoMind grounds AI responses in video transcripts, slides, and collective interaction traces. This contextual grounding can make explanations more concrete and easier for learners to interpret against the presentation materials. Consistent with prior work on grounded and interpretable AI systems [41, 67], our findings suggest that grounding also matters for learners' experience in this setting.

Importantly, EchoMind's design extends contextualization beyond material grounding to include temporal and social context. By aligning AI responses with learners' viewing position and collective difficulty patterns, the system allows AI support to reflect not only what the content says, but how others have struggled with it. This integration is consistent with the pattern observed across conditions: EchoMind was rated significantly higher than QAOnly on confusion clarification and sustained engagement, alongside a trend toward stronger future-use intentions. Within the EchoMind condition, participants likewise rated both video-grounded and behavior-aware responses as useful (Section 6.4), providing converging—though not directly comparative—evidence for the value of contextualized responses.

These findings suggest that conversational AI should not be conceptualized as a standalone tutor layered on top of learning systems.

Effective educational AI must function as a context-sensitive mediator that aligns with learning materials, temporal progress, and collective learner signals. Without such alignment, AI risks becoming an additional source of cognitive noise rather than meaningful support for learning.

## 7.3 Social Signals and the Emergence of Co-Learning

EchoMind's effect on learners' sense of co-learning highlights the importance of social visibility in asynchronous video learning. While QAOnly and VideoOnly showed no meaningful difference on this dimension, EchoMind significantly increased learners' perceptions of learning together with others. This result suggests that co-learning does not arise from AI presence alone, but from the system's ability to surface others' learning traces in ways that are temporally and contextually meaningful.

Prior work has shown that learners rely on comments, annotations, and discussion forums as social navigation cues in online learning environments [40, 43, 50, 71]. However, such cues are often detached from the precise moments when confusion occurs, limiting their ability to support in-situ sensemaking. EchoMind extends this line of work by persistently aligning collective reactions and behavioral traces with the video timeline, allowing learners to encounter social signals at the moment they are most relevant. By making others' confusion, hesitation, and understanding visible, EchoMind enables learners to situate their own comprehension within a broader social context. This social reference can help learners interpret whether a difficulty is individual or shared, consistent with prior work on social navigation and peer-linked video comments [21, 40, 71]. In this way, lightweight social signals function not merely as navigation aids, but also as interpretive supports that shape how learners read their own progress [14, 21].

Importantly, the emergence of co-learning in EchoMind is not driven by any single feature. Rather, it arises from the combined effects of emoji reactions, behavioral traces, and AI-mediated explanations that reflect collective learning states. This integrated design supports a form of collaboration that unfolds across time, without requiring synchronous interaction or explicit discussion.

## 7.4 Cognitive Fluency Without Added Workload

Although EchoMind substantially improved learners' subjective experience across multiple aspects of the learning process, these benefits were not accompanied by increased perceived workload. Across all workload measures, no significant differences were observed between conditions, and EchoMind consistently showed mean values comparable to or more favorable than those of QAOnly and VideoOnly. This pattern suggests that EchoMind can support learning processes without requiring learners to invest additional effort in interpreting or validating support. Prior research has cautioned that adding interactive or intelligent features to learning systems may increase cognitive load and disrupt learning flow [60, 75]. Our findings offer a contrasting perspective: when support is temporally aligned, contextually grounded, and socially informed, richer interactive support does not necessarily come at the cost of greater perceived workload.

In this sense, EchoMind does not add cognitive tasks; rather, it reorganizes existing cues into more accessible forms. The higher mental demand and effort reported in the QAOnly condition further illustrate this point. One plausible reading of QAOnly's descriptively higher mental demand and effort is that, without contextual grounding, learners were required to interpret and evaluate generic AI responses, which increased uncertainty and the effort required for decision-making [73]. On this reading, EchoMind's traceable, context-aligned explanations allowed learners to allocate cognitive resources to understanding rather than to verification. We describe this pattern as cognitive fluency, a state in which learners remain active while scaffolded by reliable context. In this sense, cognitive fluency in learning systems is less about minimizing interaction and more about minimizing ambiguity. When explanations are situated, traceable, and socially contextualized, learners experience learning as smoother and more coherent, even when engaging with complex content.

## 7.5 Toward a Speaker–GenAI–Audience Ecosystem: A Design Perspective

The full EchoMind architecture suggests a possible triadic ecosystem involving speakers, audiences, and generative AI, extending beyond the dyadic interaction between an individual learner and instructional content. We present this ecosystem as a design perspective rather than as a claim validated by the present study. Speaker escalation, the speaker dashboard, speaker feedback integration, dynamic knowledge-base expansion, and cross-cohort trace accumulation were not active in the evaluation. The following discussion therefore describes the interaction model that EchoMind is designed to support and motivates the longitudinal studies required to evaluate it.

Lightweight behavioral traces and expressive reactions serve as the connective tissue of this ecosystem. Rather than being treated as noise or secondary analytics, these signals function as shared learning artifacts that mediate how content is interpreted and revisited. For audiences, they provide social reference points that support reflection and self-positioning. For speakers, they reveal collective difficulty patterns that inform retrospective content refinement. For AI, they constitute essential context for generating situated and traceable explanations. Within this structure, generative AI does not operate as an autonomous tutor or a proxy for the speaker. Instead, it functions as an ecological mediator, integrating instructional materials, audience behaviors, and collective reactions into coherent, moment-aligned support. The value of AI thus lies not in replacing human interpretation, but in organizing and redistributing collective learning signals in ways that are intelligible and actionable.

The complete EchoMind architecture is intended to support the circulation of questions, reactions, behavioral traces, and speaker responses across viewers and viewing sessions. In a longitudinal deployment, these persistent signals could potentially allow later viewers to benefit from earlier interactions and provide speakers with information about recurring difficulties. However, the present study used a fixed pilot-seeded set of collective signals and did not involve speaker participation or cross-cohort accumulation. We therefore cannot determine whether the proposed triadic ecosystem

would produce cumulative interaction, collective knowledge development, or durable learning benefits. These possibilities remain hypotheses for future longitudinal evaluation.

## 7.6 Limitations and Future Work

This study has several limitations.

First, the evaluation focused on EchoMind's learner-facing experience rather than the full speaker–GenAI–audience ecosystem. The speaker dashboard and speaker-side escalation were not evaluated, and the dynamic knowledge layer was initialized with pilot traces and held constant across participants to ensure a consistent collective-information environment. Because participants completed the study asynchronously, we also could not examine escalation as an ongoing speaker–audience interaction, including potential delays between question submission, speaker handling, and learner feedback. More broadly, the study did not examine how the dynamic knowledge layer would continuously accumulate, evolve, and shape later learners' interactions in real multi-cohort settings. Thus, our findings mainly concern context-grounded conversational support, timeline-aligned collective signals, and lightweight expression for individual viewers.

Second, EchoMind was evaluated as an integrated system rather than as a set of isolated components. VideoOnly served as a non-interactive baseline, QAOnly provided basic question-answering, and EchoMind combined contextual grounding, collective signals, timeline visualizations, and expression channels. Therefore, the observed differences should be interpreted as the effect of a richer bundle of viewer-facing capabilities rather than as evidence for the contribution of any single component. More controlled comparisons or ablation studies would be needed to distinguish the respective effects of contextual grounding, collective signals, and lightweight expression mechanisms.

Third, the guided interaction requirements may have introduced an engagement-related confound into some of the subjective measures. Participants in QAOnly and EchoMind were guided to ask at least one question, while EchoMind participants were additionally guided to record at least one emoji reaction. These instructions were intended to ensure that participants experienced the interactive mechanisms available in their assigned condition. However, they also increased the amount of overt interaction in the interactive conditions by design. As a result, differences in perceived activeness, expression, or co-learning may partly reflect this additional prompted engagement, rather than the specific mechanisms through which EchoMind supported the learning experience.

Fourth, the study captured short-term experience rather than long-term learning. Participants were recruited through Prolific and watched only a five-minute learning unit, which may not reflect sustained engagement in authentic lecture settings. Although participants were instructed not to use external tools, the online setting made this difficult to enforce. Quiz performance also did not differ significantly across conditions. The present results should therefore be interpreted primarily as evidence about learners' perceived support experience during a brief viewing session, rather than as evidence for improved comprehension or durable learning outcomes. The brief, locally focused study material also means that the system's temporal retrieval design was not stress-tested: the ±30-second window may work well for highly localized concepts such as the introductory example used here, but its effectiveness for domains involving long conceptual dependencies or cumulative reasoning across larger portions of the material remains an open question.

Fifth, we did not separately measure trust in AI responses, reliance on collective cues, or susceptibility to social influence. Visible collective signals may reassure learners and help them contextualize moments of difficulty, but they may also redirect attention or encourage overreliance on frequently marked moments. Similarly, grounding responses in presentation materials may improve their perceived relevance without guaranteeing their factual correctness or appropriate trust calibration.

Future work can extend EchoMind in several directions. Studies should examine the system in more realistic learning contexts, including longer videos, repeated use, and delayed assessments, to determine whether the experiential benefits observed here translate into more durable learning outcomes. Future evaluations should also allow the dynamic knowledge layer to evolve from real learner traces over time, enabling the effects of accumulated questions, reactions, and behavioral signals to be studied across cohorts. The speaker-facing components, including escalation and the dashboard, should be evaluated in settings where speaker feedback can be meaningfully integrated, with particular attention to response delays, question prioritization, and speaker workload under concurrent demand. More controlled comparisons of system variants would help clarify the relative contributions of contextual grounding, collective signals, and lightweight expression tools. Future work should also examine how learners interpret and trust AI responses and collective cues, including whether these signals support reflection or lead to overreliance. Finally, EchoMind could be extended beyond single educational videos by linking traces across chapters or sessions and by adapting agent responses to learner history and group behavior. Similar ideas may also apply to meeting replays, professional talks, and other asynchronous knowledge-sharing contexts.

# 8 Conclusion

This paper introduced EchoMind, a system that augments recorded presentations with context-aware conversational agents grounded in both instructional materials and accumulated audience interaction signals. By integrating in-situ question answering, lightweight expressive reactions, and temporally aligned behavioral traces, EchoMind offers a design direction for transforming asynchronous video viewing from an isolated, one-off activity into a more cumulative and socially grounded learning process. Through a controlled between-subjects study, we showed that EchoMind significantly improved key aspects of learners' experience, including confusion clarification, in-the-moment expression and a sense of learning together, without increasing perceived workload. Importantly, our results indicate that these benefits do not stem from the mere presence of AI, but from how AI is situated within the video's temporal context and connected to collective learning signals.

Beyond the specific system, this work contributes a broader design perspective: treating asynchronous learning as an evolving ecosystem shaped by persistent, lightweight signals and mediated by context-aligned AI. Rather than replacing instructors or prescribing rigid interaction flows, EchoMind illustrates how AI could act as a mediator that organizes prior learners' traces into timely, interpretable support for future viewers. We believe this approach has implications not only for educational videos, but also for recorded talks, professional training, and other time-shifted knowledge-sharing contexts.

## Acknowledgments

This work was supported by the Villum Investigator grant VL-54492 by Villum Fonden. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

## References

[1] Terry Anderson, Liam Rourke, D. Randy Garrison, and Walter Archer. 2001. Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks* 5, 2 (Sept. 2001), 1–17. doi:10.24059/olj.v5i2.1875

[2] Aleksander Aristovnik, Konstantinos Karampelas, Lan Umek, and Dejan Ravšelj. 2023. Impact of the COVID-19 pandemic on online learning in higher education: a bibliometric analysis. In *Frontiers in Education*, Vol. 8. Frontiers Media SA, 1225834. doi:10.3389/feduc.2023.1225834

[3] Benjamin Artz, Marianne Johnson, Denise Robson, and Sarinda Siemers. 2022. Live or lecture capture: Evidence from a classroom random control trial. *International Review of Economics Education* 40 (2022), 100240. doi:10.1016/j.iree.2022.100240

[4] Christopher Aylwin. 2019. Faculty and Student Interaction in an Online Master's Course: Survey and Content Analysis. *JMIR Medical Education* 5, 1 (2019), e10464. doi:10.2196/10464

[5] Qiyu Bai, Qi Dan, Zhe Mu, and Maokun Yang. 2019. A Systematic Review of Emoji: Current Research and Future Perspectives. *Frontiers in Psychology* Volume 10 - 2019 (2019). doi:10.3389/fpsyg.2019.02221

[6] Seyyed Kazem Banihashem, Omid Noroozi, Stan Van Ginkel, Leah P Macfadyen, and Harm JA Biemans. 2022. A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review* 37 (2022), 100489. doi:10.1016/j.edurev.2022.100489

[7] Anna Bodonhelyi, Enkeleda Thaqi, Süleyman Özdel, Efe Bozkir, and Enkelejda Kasneci. 2025. From Passive Watching to Active Learning: Empowering Proactive Participation in Digital Classrooms with AI Video Assistant. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 16:1–16:21. doi:10.1145/3706598.3713513

[8] Christopher G. Brinton, Swapna Buccapatnam, Mung Chiang, and H. Vincent Poor. 2016. Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance. *Trans. Sig. Proc.* 64, 14. doi:10.1109/TSP.2016.2546228

[9] John Brooke. 1996. SUS: A "Quick and Dirty" Usability Scale. In *Usability Evaluation in Industry*, Patrick W. Jordan, Bruce Thomas, and Bernard A. Weerdmeester (Eds.). Taylor & Francis, London, UK, 189–194.

[10] Jennifer Campbell, Alison Gibbs, Hedieh Najafi, and Cody Severinski. 2014. A comparison of learner intent and behaviour in live and archived MOOCs. *International Review of Research in Open and Distributed Learning* 15, 5 (2014), 235–262. doi:10.19173/irrodl.v15i5.1854

[11] Jennifer Campbell, Diane Horton, Michelle Craig, and Paul Gries. 2014. Evaluating an inverted CS1. In *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. ACM, New York, NY, USA, 307–312. doi:10.1145/2538862.2538943

[12] Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. 2021. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence* 2 (2021), 100027. doi:10.1016/j.caeai.2021.100027

[13] Si Chen, Jason Situ, Haocong Cheng, Desirée Kirst, and Yun Huang. 2023. MirrorUs: Mirroring Peers' Affective Cues to Promote Learner's Meta-Cognition in Video-based Learning. *Proceedings of the ACM on Human-Computer Interaction* CSCW2 (2023). doi:10.1145/3610079

[14] Yue Chen, Qin Gao, and Ge Gao. 2022. Timeline-Anchored Comments in Video-Based Learning: The Impact of Visual Layout and Content Depth. *International Journal of Human–Computer Interaction* 38, 9 (2022), 868–883. doi:10.1080/10447318.2021.1976505

[15] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann, and Robert Glaser. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13, 2 (1989), 145–182. doi:10.1207/s15516709cog1302_1

[16] Michelene T. H. Chi and Ruth Wylie. 2014. The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist* 49, 4 (2014), 219–243. doi:10.1080/00461520.2014.965823

[17] Dan Davis, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2018. Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education* 125 (2018), 327–344. doi:10.1016/j.compedu.2018.05.019

[18] Vanessa P. Dennen, A. Aubteen Darabi, and Linda J. Smith. 2007. Instructor-Learner Interaction in Online Courses: The relative perceived importance of particular instructor actions on performance and satisfaction. *Distance Education* 28, 1 (2007), 65–79. doi:10.1080/01587910701305319

[19] Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2 (2012), 145–157. doi:10.1016/j.learninstruc.2011.10.001

[20] Sidney D'Mello and Art Graesser. 2013. AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 4, Article 23 (2013), 39 pages. doi:10.1145/2395123.2395128

[21] Thomas Erickson and Wendy A. Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction* 7, 1 (2000), 59–83. doi:10.1145/344949.345004

[22] Jingchao Fang, Jeongeon Park, Juho Kim, and Hao-Chuan Wang. 2024. EduLive: Re-Creating Cues for Instructor-Learners Interaction in Educational Live Streams with Learners' Transcript-Based Annotations. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–33. doi:10.1145/3686960

[23] C. Ailie Fraser, Tricia J. Ngoon, Mira Dontcheva, and Scott Klemmer. 2019. RePlay: contextually presenting learning videos across software applications. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3290605.3300527

[24] D. Randy Garrison, Terry Anderson, and Walter Archer. 2000. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education* 2, 2-3 (2000), 87–105. doi:10.1016/S1096-7516(00)00016-6

[25] Pratik Ghosh and Sean Rintel. 2025. YES AND: A Generative AI Multi-Agent Framework for Enhancing Diversity of Thought in Individual Ideation for Problem-Solving Through Confidence-Based Agent Turn-Taking. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM, New York, NY, USA. doi:10.1145/3706599.3720142

[26] Michail N. Giannakos, Konstantinos Chorianopoulos, and Nikos Chrisochoides. 2015. Making Sense of Video Analytics: Lessons Learned from Clickstream Interactions, Attitudes, and Learning Outcome in a Video-Assisted Course. *International Review of Research in Open and Distributed Learning* 16, 1 (2015), 260–283. doi:10.19173/irrodl.v16i1.1976

[27] Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A spatially anchored census of student confusion for online lecture videos. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1555–1564. doi:10.1145/2702123.2702304

[28] Arthur C Graesser, Kurt VanLehn, Carolyn P Rosé, Pamela W Jordan, and Derek Harter. 2001. Intelligent tutoring systems with conversational dialogue. *AI magazine* 22, 4 (2001), 39–52. doi:10.1609/aimag.v22i4.1591

[29] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: an investigation of contextual video assistance for functionality understanding. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1515–1524. doi:10.1145/1753326.1753552

[30] Philip J. Guo, Juho Kim, and Rob Rubin. 2014. How video production affects student engagement: an empirical study of MOOC videos. In *Proceedings of the ACM Conference on Learning @ Scale Conference*. ACM, New York, NY, USA, 41–50. doi:10.1145/2556325.2566239

[31] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183. doi:10.1016/S0166-4115(08)62386-9

[32] Atsusi Hirumi. 2009. A Framework for Analyzing, Designing, and Sequencing Planned Elearning Interactions. In *The Perfect Online Course: Best Practices for Designing and Teaching*, Anymir Orellana, Terry L. Hudgins, and Michael Simonson (Eds.). Emerald Publishing Limited. doi:10.1108/978-1-60752-205-820251013

[33] Zipeng Ji, Pengcheng An, and Jian Zhao. 2025. ClassComet: Exploring and Designing AI-generated Danmaku in Educational Videos to Enhance Online Learning. In *Proceedings of the ACM Conference on Designing Interactive Systems*. ACM, New York, NY, USA, 552–575. doi:10.1145/3715336.3735762

[34] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221* (2022).

[35] Eun Jung Kim and Mi Jeong Kim. 2024. Students' Evaluation of Live Online Lectures Versus Pre-Recorded Video Lectures During COVID-19: Focusing on Students' Study Behaviors. *Sage Open* 14, 4 (2024). doi:10.1177/21582440241305325

[36] Juho Kim, Philip J. Guo, Daniel T. Seaton, Piotr Mitros, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the First ACM Conference on Learning @ Scale Conference*. ACM, New York, NY, USA, 31–40. doi:10.1145/2556325.2566237

[37] Nataliya Kosmyna and Pattie Maes. 2019. AttentivU: An EEG-Based Closed-Loop Biofeedback System for Real-Time Monitoring and Improvement of Engagement for Personalized Learning. *Sensors* (2019). doi:10.3390/s19235200

[38] Shelbi L Kuhlmann, Robert Plumley, Zoe Evans, Matthew L Bernacki, Jeffrey A Greene, Kelly A Hogan, Michael Berro, Kathleen Gates, and Abigail Panter. 2024. Students' active cognitive engagement with instructional videos predicts STEM learning. *Computers & Education* 216 (2024). doi:10.1016/j.compedu.2024.105050

[39] Hyowon Lee, Mingming Liu, Michael Scriney, and Alan F. Smeaton. 2022. Playback-centric visualizations of video usage using weighted interactions to guide where to watch in an educational context. *Frontiers in Education* 7 (2022), 1–16. doi:10.3389/feduc.2022.733646

[40] Yi-Chieh Lee, Wen-Chieh Lin, Fu-Yin Cherng, Hao-Chuan Wang, Ching-Ying Sung, and Jung-Tai King. 2015. Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 689–698. doi:10.1145/2702123.2702349

[41] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 16 pages. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[42] Nan Li, Łukasz Kidziński, Patrick Jermann, and Pierre Dillenbourg. 2015. MOOC Video Interaction Patterns: What Do They Tell Us?. In *Design for Teaching and Learning in a Networked World*. Springer International Publishing, Cham, 197–210. doi:10.1007/978-3-319-24258-3_15

[43] Jun Rong Nigel Lim, Sonny Rosenthal, Ye Jun Marcius Sim, Zhao-Yi Lim, and Kai Rong Oh. 2021. Making online learning more satisfying: The effects of online-learning self-efficacy, social presence and content structure. *Technology, Pedagogy and Education* 30, 4 (2021), 543–556. doi:10.1080/1475939X.2021.1934102

[44] Qianyu Liu, Xinran Li, Xiaocong Du, and Quan Li. 2025. TSConnect: An Enhanced MOOC Platform for Bridging Communication Gaps Between Instructors and Students in Light of the Curse of Knowledge. In *Proceedings of the ACM Conference on Intelligent User Interface*. ACM, New York, NY, USA, 1335–1353. doi:10.1145/3708359.3712108

[45] Shuai Ma, Taichang Zhou, Fei Nie, and Xiaojuan Ma. 2022. Glancee: An Adaptable System for Instructors to Grasp Student Learning Status in Synchronous Online Classes. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 313, 25 pages. doi:10.1145/3491102.3517482

[46] Pratyusha Maiti and Ashok Goel. 2025. Can an AI Partner Empower Learners to Ask Critical Questions?. In *Proceedings of the ACM Conference on Intelligent User Interface*. ACM, New York, NY, USA, 314–324. doi:10.1145/3708359.3712134

[47] Martin Merkt, Anne Ballmann, Julia Felfeli, and Stephan Schwan. 2018. Pauses in educational videos: Testing the transience explanation against the structuring explanation. *Computers in Human Behavior* 89 (2018), 399–410. doi:10.1016/j.chb.2018.01.013

[48] Toni-Jan Keith Palma Monserrat, Shengdong Zhao, Kevin McGee, and Anshul Vikram Pandey. 2013. NoteVideo: Facilitating navigation of blackboard-style lecture videos. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1139–1148. doi:10.1145/2470654.2466147

[49] Michael G. Moore. 1989. Editorial: Three types of interaction. *American Journal of Distance Education* 3, 2 (1989), 1–7. doi:10.1080/08923648909526659

[50] Esteban Morales, Jeremiah H Kalir, Alice Fleerackers, and Juan Pablo Alperin. 2022. Using social annotation to construct knowledge with others: A case study across undergraduate courses. *F1000Research* 11 (2022), 235. doi:10.12688/f1000research.109525.2

[51] Taichi Murakami, Kazuyuki Fujita, Kotaro Hara, Kazuki Takashima, and Yoshifumi Kitamura. 2024. SwapVid: Integrating Video Viewing and Document Exploration with Direct Manipulation. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1035:1–1035:13. doi:10.1145/3613904.3642515

[52] Prasanth Murali, Javier Hernandez, Daniel McDuff, Kael Rowan, Jina Suh, and Mary Czerwinski. 2021. AffectiveSpotlight: Facilitating the Communication of Affective Responses from Audience Members during Online Presentations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA. doi:10.1145/3411764.3445235

[53] Cuong Nguyen and Feng Liu. 2015. Making software tutorial video responsive. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1565–1568. doi:10.1145/2702123.2702209

[54] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* 54, 4 (2022), 1643–1662. doi:10.3758/s13428-021-01694-3

[55] Boris Ruf and Marcin Detyniecki. 2024. The "Huh?" Button: Improving Understanding in Educational Videos with Large Language Models. In *Proceedings of the IEEE International Symposium on Multimedia*. IEEE Computer Society, Los Alamitos, CA, USA, 285–289. doi:10.1109/ISM63611.2024.00064

[56] Anne-Dominique Salamin, David Russo, and Danièle Rueger. 2023. ChatGPT, an excellent liar: how conversational agent hallucinations impact learning and teaching. In *Proceedings of the 7th International Conference on Teaching, Learning and Education*. doi:10.33422/6th.iacetl.2023.11.100

[57] Hyungyu Shin, Eun-Young Ko, Joseph Jay Williams, and Juho Kim. 2018. Understanding the effect of in-video prompting on learners and instructors. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3173574.3173893

[58] Muhtasim Ibteda Shochcho, Mohammad Ashfaq Ur Rahman, Shadman Rohan, Ashraful Islam, Hasnain Heickal, AKM Mahbubur Rahman, M. Ashraful Amin, and Amin Ahsan Ali. 2025. Improving User Engagement and Learning Outcomes in LLM-Based Python Tutor: A Study of PACE. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3706599.3720240

[59] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, Carolyn Rose and George Siemens (Eds.). Association for Computational Linguistics, Doha, Qatar, 3–14. doi:10.3115/v1/W14-4102

[60] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76. doi:10.1016/B978-0-12-387691-1.00002-8

[61] Chek Tien Tan, Indriyati Atmosukarto, Budianto Tandianus, Songjia Shen, and Steven Wong. 2025. Exploring the Impact of Avatar Representations in AI Chatbot Tutors on Learning Experiences. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1072:1–1072:12. doi:10.1145/3706598.3713456

[62] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2023. Scripted vicarious dialogues: Educational video augmentation method for increasing isolated students' engagement. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–25. doi:10.1145/3544548.3581153

[63] Loren Terveen and Will Hill. 1998. Evaluating Emergent Collaboration on the Web. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. ACM, New York, NY, USA, 355–362. doi:10.1145/289444.289510

[64] Susanne Voelkel, Andy Bates, Terry Gleave, Carl Larsen, Elliott J. Stollar, Gemma Wattret, and Luciane V. Mello. 2023. Lecture capture affects student learning behaviour. *FEBS Open Bio* 13, 2 (2023), 217–232. doi:10.1002/2211-5463.13548

[65] Tianjia Wang, Tong Wu, Huayi Liu, Chris Brown, and Yan Chen. 2025. Generative Co-Learners: Enhancing Cognitive and Social Presence of Students in Asynchronous Learning with Generative AI. *Proceedings of the ACM on Human-Computer Interaction* (2025). doi:10.1145/3701198

[66] Xinyu Jessica Wang, Christine P Lee, and Bilge Mutlu. 2025. LearnMate: Enhancing Online Education with LLM-Powered Personalized Learning Plans and Support. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–10. doi:10.1145/3706599.3719857

[67] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng (Polo) Chau. 2021. CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1396–1406. doi:10.1109/TVCG.2020.3030418

[68] Jiun-Yu Wu, Chen-Hsuan Liao, Chin-Chung Tsai, and Oi-Man Kwok. 2024. Using learning analytics with temporal modeling to uncover the interplay of before-class video viewing engagement, motivation, and performance in an active learning context. *Computers & Education* 212 (2024), 104975. doi:10.1016/j.compedu.2023.104975

[69] Jiarui Xie and Ana-Paula Correia. 2024. The effects of instructor participation in asynchronous online discussions on student performance: A systematic review. *British Journal of Educational Technology* 55, 1 (2024), 71–89. doi:10.1111/bjet.13350

[70] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (2024), 90–112. doi:10.1111/bjet.13370

[71] Matin Yarmand, Dongwook Yoon, Samuel Dodson, Ido Roll, and Sidney S Fels. 2019. "Can you believe [1: 21]?!" Content and Time-Based Reference Patterns in Video Comments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3290605.3300719

[72] Meehyun Yoon, Jungeun Lee, and Il-Hyun Jo. 2021. Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *The Internet and Higher Education* 50 (2021). doi:10.1016/j.iheduc.2021.100806

[73] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28. doi:10.1186/s40561-024-00316-7

[74] Yuanhao Zhang, Yumeng Wang, Xiyuan Wang, Changyang He, Chenliang Huang, and Xiaojuan Ma. 2025. CoKnowledge: Supporting Assimilation of Time-synced Collective Knowledge in Online Science Videos. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, Article 632, 22 pages. doi:10.1145/3706598.3713682

[75] Zihan Zhang, Black Sun, and Pengcheng An. 2025. Breaking Barriers or Building Dependency? Exploring Team-LLM Collaboration in AI-infused Classroom Debate. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–19. doi:10.1145/3706598.3713853

[76] Yihao Zhu, Zhoutong Ye, Yichen Yuan, Wenxuan Tang, Chun Yu, and Yuanchun Shi. 2025. AutoPBL: An LLM-powered Platform to Guide and Support Individual Learners Through Self Project-based Learning. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 584:1–584:26. doi:10.1145/3706598.3714261

# A Appendix

## A.1 DESIGNED QUESTIONS IN QUESTIONNAIRES

The following describes the detailed questions in post-task questionnaires.

**Table 2: Post-task questionnaire for participants on 7-point Likert scales (agreement items: 1: Strongly disagree, 7: Strongly agree; workload items: 1: Not at all, 7: Extremely).**

| Category | Question |
|---|---|
| **Usability** | 1) The system was easy to use.<br>2) The system supported my learning process effectively.<br>3) I am satisfied with the platform.<br>4) I would like to use a system like this when watching educational videos in the future. |
| **Workload** | 1) How mentally demanding was the task?<br>2) How successful do you think you were in completing the task?<br>3) How much effort did you have to invest to perform well?<br>4) How discouraged, stressed, or frustrated did you feel? |
| **Learning experience** | 1) The system helped me clarify points I was confused about.<br>2) The system helped me stay engaged during the video.<br>3) The system encouraged me to reflect and think actively during learning.<br>4) The system encouraged me to express questions or emotions at the moment they occurred.<br>5) I felt like I was learning together with others, not just by myself.<br>6) Overall, my learning process felt more active rather than passive. |
| **Perceptions of the system** | 1) Seeing others' reactions and using emoji reactions helped me reflect on my understanding rather than passively watching.<br>2) The Cognitive Curve helped me notice where learners tended to struggle.<br>3) The Reaction Intensity chart helped me understand the overall learning status of the audience.<br>4) AI responses grounded in the video materials helped me understand the content better.<br>5) AI responses that considered user behavior were useful.<br>6) Using emoji reactions made me feel like I was actively participating, not just watching the video. |