CoRead: Using Implicit Behavioural Signals for Mixed-Initiative AI Support in Collaborative Academic Reading

Sakil Sarker
Ontario Tech University
Oshawa, Ontario, Canada
sakil.sarker@ontariotechu.net

Yasaman A.Basti
Computer Science
Ontario Tech University
Oshawa, Ontario, Canada
Yasaman.akhoundzadehbasti@ontariotechu.net

Heidar Davoudi
Computer Science
Ontario Tech University
Oshawa, Ontario, Canada
heidar.davoudi@ontariotechu.ca

Ali Neshati
Computer Science
Ontario Tech University
Oshawa, Ontario, Canada
ali.neshati@ontariotechu.ca

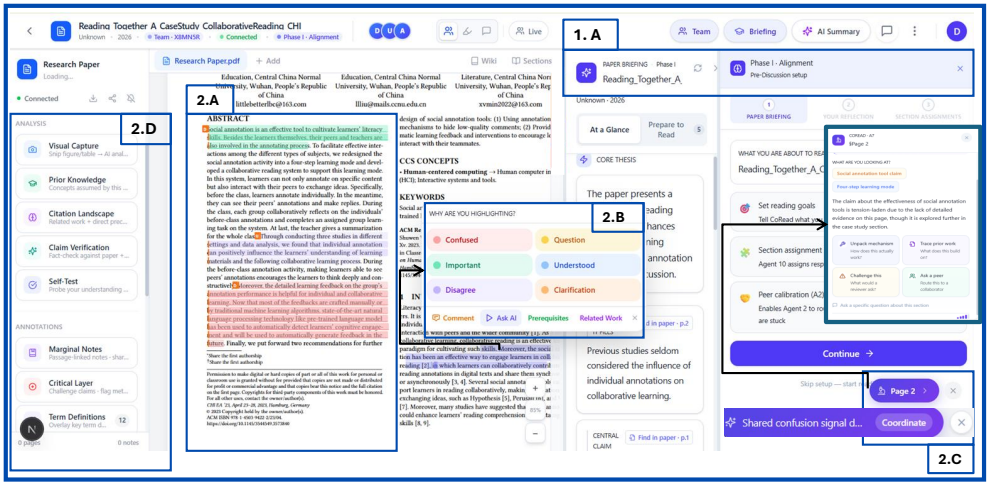


Figure 1: CoRead, a mixed-initiative collaborative reading interface that detects silent confusion and routes help without requiring readers to signal struggle publicly. Phase I (1.A), structured briefing and section assignment before the session. Phase II (2.A–2.D), confusion detected from implicit signals; annotation reason tags (2.B) and a proactive page pill (2.C) surface help; analysis tools remain available (2.D). Phase III (3.A), session synthesised into a Shared Understanding document.

Abstract

Collaborative academic reading depends on coordination that existing tools do not provide. Readers need to know when peers are confused, who can help, and when AI assistance would add value rather than interrupt. Two interaction problems remain unresolved: *when* AI should intervene (grounded in sustained behavioural signals rather than isolated events or explicit requests) and *how* assistance should be orchestrated across individuals, peers, and the group. We present CoRead, a mixed-initiative collaborative reading interface addressing both. A formative study ($N = 12$) grounded five design requirements informing a coordination architecture that routes help across pre-reading alignment, in-session support, and post-session synthesis, abstaining when intervention would disrupt more than it helps. A Wizard-of-Oz feasibility evaluation ($N = 12$) showed 81.1% acceptance of implicit AI interventions and 53.8% acceptance of peer-routing offers, SUS = 73.3 (95% CI [69.7, 76.9]), NASA-TLX = 47.0, and automated cosine-similarity grounding checks passing 19 of 21 AI-generated explanations.

This work is licensed under a Creative Commons Attribution 4.0 International License.
ICMI '26, Napoli, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2318-6/26/10
https://doi.org/10.1145/3776574.3831179

CCS Concepts

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

Keywords

Collaborative reading; Proactive AI assistance; Human-AI collaboration; Mixed-initiative systems; Academic sensemaking; Behavioural sensing

ACM Reference Format:
Sakil Sarker, Yasaman A.Basti, Heidar Davoudi, and Ali Neshati. 2026. CoRead: Using Implicit Behavioural Signals for Mixed-Initiative AI Support in Collaborative Academic Reading. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26), October 05–09, 2026, Napoli, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3776574.3831179

1 Introduction

Collaborative academic reading is central to graduate research. Groups read together to pool expertise, surface interpretive disagreements, and build shared understanding of complex material [37]. In practice, coordination rarely works as intended. When a reader stalls on a methods section, scrolls back, and re-reads the same equation three times, nothing reaches their collaborators; by the time confusion surfaces in discussion, the moment for targeted help has passed [25]. Existing tools leave two interaction problems unresolved [7, 14, 31].

The first concerns *timing*. When should AI intervene, and how should that decision be grounded in sustained behavioural signals rather than isolated events or explicit requests? Annotation systems make reading visible, but only what readers choose to share [26, 36]; AI reading tools respond well to direct questions but wait to be asked [17]. Both assume readers recognise and articulate their own confusion, yet the formative evidence in Section 3 shows confusion accumulates silently through behaviour long before it is voiced.

The second concerns *orchestration*. Once confusion is detected, how should assistance be routed across individuals, peers, and the group? Multimodal behavioural sensing shows reading behaviour already reveals when a reader is confused [9, 24]. Stiber et al. [35] demonstrated these signals can trigger AI help in physical assembly tasks, but only for individual users. No prior work has used them to decide *who else* in a collaborative group should be involved, whether to help the stuck reader directly, route them to a knowledgeable peer, or surface the difficulty to the whole group. **CoRead addresses both as a unified coordination challenge.**

To our knowledge, CoRead is the first system to combine continuous multimodal confusion detection with active coordination across individuals, peers, and groups during collaborative academic reading. A formative study ($N = 12$, REB File No. 18326) grounded the design in five findings; a feasibility evaluation ($N = 12$, REB File No. 19432) assessed it in practice. This paper makes three contributions.

- **A formative account** of how confusion accumulates silently during collaborative academic reading, from interviews with 12 researchers, producing five design requirements and a phase-aware framework spanning pre-reading alignment, in-session orchestration, and post-session stabilization.
- **CoRead**, a working system instantiating this framework through a section-indexed difficulty score D_s computed from six multimodal signals, stability-controlled routing under persistence constraints, and a 10-agent coordination architecture with section-constrained assistance that abstains rather than hallucinating.
- **Empirical evidence** from a 12-participant Wizard-of-Oz feasibility evaluation, with 81.1% implicit intervention acceptance, 53.8% peer-routing acceptance, SUS = 73.3 (95% CI [69.7, 76.9]), NASA-TLX = 47.0, and automated grounding checks passing 19 of 21 invocations.

2 Related Work

Collaborative reading and annotation. Marshall and Brush established that making reader traces visible scaffolds shared interpretation and group coordination [25, 26]. LiquidText extended this with flexible passage-linking and spatial organization during active reading [36]. Social annotation platforms such as Perusall and NB showed that instructor visibility into reader confusion improves comprehension and discussion quality [41], and research on collaborative sensemaking confirms that shared external representations help groups converge [31]. All of these systems depend on readers actively choosing to externalise confusion. A reader who is struggling but producing no annotations generates no signal for collaborators, and no annotation system monitors whether grounding is occurring. This distinction also means that collaborative annotation tools are not a like-for-like control for CoRead: annotation systems externalise difficulty that readers have already chosen to mark, whereas CoRead attempts to respond before any explicit signal is produced.

Coordination and workspace awareness. Foundational CSCW research on workspace awareness establishes that collaborators need continuous access to information about each other's activities, intentions, and states [10, 14]; when awareness breaks down, groups lose alignment and confusion accumulates without surfacing into discussion [14]. Schmidt and Simone argue that effective cooperative work depends on coordination mechanisms that actively maintain shared awareness [34]. No existing system acts on this awareness to restore alignment when a breakdown is detected.

Behavioural sensing for confusion detection. D'Mello and Graesser documented reliable multimodal confusion signals during complex learning (rereading loops, response latency shifts, irregular interaction patterns) strong enough for automatic detection [9]. WEDAR showed webcam gaze and body language identify six attention-related behaviours during e-reading with 75% accuracy [24]. Stiber et al. connected behavioural sensing to AI-triggered assistance in physical assembly tasks [35]. Guo et al. combined eye-tracking with LLM reasoning for collaborative programming, showing that multi-user signals analyzed jointly yield more contextually relevant AI help [13]. None of this work addresses what should happen in a collaborative group *once* confusion is detected, whether to help directly, involve a knowledgeable peer, or surface the difficulty to the group.

Mixed-initiative and proactive AI assistance. Horvitz established that a proactive action is worth taking only when its benefit outweighs the cost of interruption, and that behavioural context must inform that judgement [18, 19]. Fogarty et al. and Iqbal and Bailey confirmed that poorly timed interventions impose measurable costs, and that deferring to task breakpoints reduces disruption [12, 21]. Amershi et al. codified these findings into design guidelines emphasizing transparency, user control, and graceful

degradation [1]. AI reading tools such as ScholarPhi [17] operationalise some of these principles but remain reactive. Zhang et al. and Flathmann et al. showed AI teammates must model the team's state [11, 40], but proactive intervention across groups with varying expertise and confusion states has not been addressed.

No existing system senses sustained confusion from behaviour and routes the right form of help across individuals, peers, and groups — CoRead addresses this gap.

3 Formative Study

Before building CoRead, we needed a grounding in researchers' collaborative reading experience to answer three research questions (RQ) that existing literature leaves open.

- **RQ1 (signals).** How does reading behaviour provide evidence of sustained difficulty before it surfaces verbally, and how can such evidence inform assistance routing?
- **RQ2 (intervention).** What timing, form, and degree of control makes AI assistance feel helpful rather than intrusive?
- **RQ3 (structure).** What collaboration structure (roles, phases, shared artifacts) makes peer coordination work in practice?

3.1 Participants and Procedure

We recruited 12 participants through purposive sampling, researchers who regularly read papers in lab reading groups, journal clubs, or thesis group sessions in CS or related disciplines. The cohort comprised 10 master's-level researchers and 2 faculty (7 women, 5 men; age 20–45; P1–P12), spanning CS, NLP, HCI, and adjacent fields, all with at least one year of collaborative reading experience. Each semi-structured interview lasted 45–60 minutes (REB File No. 18326). Interviews covered the three RQs and were audio-recorded and transcribed.

3.2 Analysis

Transcripts were analyzed using codebook thematic analysis [3, 4]. Two researchers collaboratively developed an initial codebook from four transcripts (33%), producing 36 codes, then independently coded a randomly selected 20% subsample. Cohen's $\kappa = 0.81$ ($p < .001$) indicated strong agreement [27]. Disagreements were resolved through discussion. Remaining transcripts were coded by the first author using the finalised codebook, with ambiguous cases flagged for joint review.

3.3 Findings

Analysis yielded five findings (F1 addresses RQ1; F2 and F4 address RQ2; F3 and F5 address RQ3). Table 1 maps each to its design requirement.

F1 (confusion is behavioural before verbal), RQ1. All 12 described confusion building through rereading loops and annotation clusters around equations ($n = 12$), figures ($n = 10$), and unfamiliar terms ($n = 9$) before anyone spoke aloud. Three described individual difficulty escalating into group-level stalls. As P1 put it, *"Several of us get stuck on the same part… we reread the section and try to figure out what the equation means."* P8 asked for a system that could "silently observe reading patterns such as rereading or reading slowly."

Table 1: Five formative findings ($N = 12$) and the design requirements they motivated. #P = number of participants expressing the finding.

Finding (#P)	Design requirement
F1. Confusion is behavioural before verbal (12)	Monitor signals continuously at section granularity; detect sustained difficulty before it surfaces.
F2. Timely, controlled AI help (12)	Gate behind persistence window; non-modal and dismissible; opt-out at every stage; depth-levelled on acceptance.
F3. Collaboration needs orchestration (12)	Assign roles pre-session; route to peers in-session; synthesise shared artifacts post-session.
F4. Output density must be managed (10)	Deduplicate annotations; throttle AI outputs; passive synthesis preferred over real-time broadcast.
F5. AI must be traceable to source (9)	Section-constrained generation; abstain when grounding fails; source badges on peer claims.

F2 (timely, controlled AI help), RQ2. All 12 accepted AI assistance only after *sustained* difficulty, not a single reread. P10 noted, *"If I am stuck on something for a while, then AI could step in."* Form mattered equally. P10 required "something on the side… a small icon or even a light border"; P2 rejected immediate pop-ups ("it's like browsing a website and an ad pops up"). Depth was a further condition. P1 asked AI to "explain like a beginner first, then introduce previous methods"; P12 wanted a "simpler idea first; once I get that, then I go deeper."

F3 (collaboration needs orchestration), RQ3. All 12 wanted AI to synthesise collaborative activity into structured summaries rather than surface raw logs. Eight wanted AI to propose discussion points. P1 suggested, *"Find out three or five points of discussion… pose it to the team."* Six preferred peer explanation over AI when a knowledgeable colleague was available, and six described pre-reading role clarity as essential for comfortable peer interaction. This motivates role assignment *before* sessions so peer routing has a principled basis.

F4 (output density must be managed), RQ2. Seven warned that dense collaborative environments produce overload. P12 warned, *"If five users are all interacting in different ways… there could easily be hundreds of comments and highlights… that could get overwhelming fast."* Ten preferred passive modes, reading first then accessing synthesised summaries.

F5 (AI must be traceable to source), RQ3. Nine described trust as depending on traceability, not accuracy alone. P9 stressed, *"Trust is crucial… AI hallucinates… I would hope that it is backed up."* P3 raised the same concern for peer claims. Participants preferred options that highlighted relevant passages besides explanations.

Design tensions. Two participants rejected implicit assistance entirely, motivating a full opt-out mode; one questioned collaborative reading altogether, motivating a solo mode. Several warned that CoRead's own outputs could become overload sources, motivating suppression and cooldown mechanisms.

4 The CoRead System

CoRead operationalises the five findings as a web-based collaborative reading environment organised into three coordination phases. *Phase I* structures who reads what before the session; *Phase II* detects silent struggle and routes help; *Phase III* synthesises what the group understood after the session. No prior tool integrates all three phases under a shared behavioural signal model (Figure 1).[1]

4.1 Signal Model and Routing Architecture

All three phases share a common signal layer. Each section s maintains a per-section difficulty score $D_s \in [0, 1]$, engagement stability E_s (forward-scroll distance relative to regression frequency over 30 s, distinguishing deep reading from stagnation), and section responsibility strength R_s (keyword overlap between each reader's Phase I reflection and the section heading, normalised to $[0, 1]$). D_s is the only component with a full formal model; E_s and R_s operate as heuristic auxiliary signals that suppress false positives and select peers, respectively.

D_s is a composite heuristic score whose weights and firing thresholds are hand-tuned estimates rather than empirically optimised parameters; we return to empirical validation in Section 6. It is computed every 2 s by A1 from six multimodal signals across three modalities.

$$D_s = \min\left(1,\ \sum_{i \in \mathcal{B}} w_i \cdot s_i \ + \ \sum_{j \in \mathcal{X}} w_j \cdot s_j + \lambda \cdot t\right) \tag{1}$$

Here $\mathcal{B}$ is the implicit behavioural tier (always active), $\mathcal{X}$ the explicit annotation tier (amplification only), and $\lambda = -0.012\,\mathrm{s}^{-1}$ a decay term returning D_s to baseline within 60 s. The two-tier design ensures detection functions for readers who never annotate while rewarding annotation with faster assistance. If WebGazer is unavailable, the frozen-stare weight is zeroed and D_s uses the remaining five signals. This fallback is not evidence of modality-level performance but an implementation property.

F2 established that readers want help only after *sustained* difficulty. CoRead implements *stability-controlled routing*, assistance dispatched only when D_s exceeds τ_{fire} continuously for persistence window Δt with low E_s.

$$D_s(t) > \tau_{\text{fire}} \ \wedge \ E_s(t) < \tau_e \quad \forall\, t \in [t_0,\, t_0 + \Delta t] \tag{2}$$

Thresholds were set through iterative piloting ($n = 3$ sessions, 4 participants) using two criteria, false-positive rate below 20% and acceptance rate above 70%. $\tau_e = 0.35$ [33]; $\tau_{\text{fire}} = 0.72$ [21]; $\tau_{\text{high}} = 0.88$ [9]. Signal weights in Table 2 follow the relative predictive strength reported in [9, 33] and were refined during piloting. Per-reader adaptive thresholds are discussed in Section 6.

Routing escalates across four levels.

(1) **Silence** ($E_s \geq \tau_e$). Forward progress dominates; interventions suppressed (F4).
(2) **AI notification** ($\tau_{\text{fire}} \leq D_s < \tau_{\text{high}}$). A7 fires a three-stage notification; A6 generates grounded explanation on acceptance (F2).
(3) **Peer routing** ($D_s \geq \tau_{\text{high}}$). A2 routes to a knowledgeable peer via R_s (F3).
(4) **Group study mode** ($|\{u \mid D_{s,u} > \tau_{\text{fire}}\}| \geq 2$). A3 seeds collective discussion anchored to the shared struggle section (F3).

A 90 s per-section cooldown prevents oscillatory prompting, matching Iqbal and Bailey's 64–87 s resumption lag [21]. All routing decisions are logged. Table 3 summarises all ten agents.

Table 2: Multimodal signals and weights for D_s. Implicit tier ($\mathcal{B}$, always active) forms the primary detection layer; explicit tier ($\mathcal{X}$) amplifies D_s when readers tag confusion. Weights follow [9, 33].

Signal	w_i	Modality
Implicit, always active ($\mathcal{B}$)		
Rereading loop (scroll-back to same passage)	+0.28	Scroll
Scroll regression (reverse scroll in section)	+0.22	Scroll
Frozen stare (>2 s gaze dwell, same line)	+0.18	Gaze
Mouse jitter (erratic pointer pattern)	+0.10	Pointer
Extended dwell time on same section	+0.08/s	Scroll
Section content density modifier	+0.06	Document
Score decay (λ)	−0.012/s	n/a
Explicit, amplification tier ($\mathcal{X}$)		
Stuck marker (*"I am stuck"* click)	+0.30	Annotation
Confusion highlight	+0.25	Annotation

Table 3: CoRead's ten agents by phase, role, and formative finding addressed.

Agent	Role	F
Phase I, Pre-Reading Alignment		
A6 Content Comprehension	Paper Briefing; grounded pre-reading orientation	F1,F5
A10 Role Allocator	Assigns sections with one-sentence rationale	F3
Phase II, In-Session Orchestration		
A1 Struggle Detector	Computes D_s every 2 s	F1
A2 Collab. Orchestrator	Routes to peer at $D_s \geq \tau_{\text{high}}$	F3
A3 Discussion Facilitator	Section-anchored prompts; seeds peer/group chat	F3
A4 Annotation Analyser	Deduplicates; surfaces divergence	F4
A6 Content Comprehension	Depth-levelled Help Card	F2,F5
A7 Implicit Assistance	Three-stage notification	F2
A8 Fact Checker	Source badges on peer claims	F5
A9 Related Work Analyser	Citations for unfamiliar terms	F1
Phase III, Post-Session Stabilisation		
A5 Storyboard Curator	Shared Understanding document	F3,F5

4.2 Pre-Reading Alignment (Phase I)

Phase I provides structured pre-reading alignment so peer routing has a principled basis rather than guesswork (F3).

Paper Briefing (A6). A6 generates a Paper Briefing covering the paper's core thesis, the gap it fills, central claim, validation, and a self-triage note; every field links to its PDF source (F5). A *Before*

[1] Stack, Next.js 15, Apryse WebViewer (PDF, XFDF annotation), Socket.IO 4, Prisma + Firebase, GPT-4o [28] at temperature 0.2, WebGazer.js [29] (gaze, optional), pdfHeadingExtractor.

You Read panel lists prerequisite concepts and prior papers (F1). P1 asked for "a cheat sheet from other papers."

Reading Purpose Reflection (A10). Each reader submits a short reflection stating reading purpose and prior familiarity, populating R_s for every section so peer routing is based on declared expertise rather than inference (F3).

Section Assignment (A10). A10 assigns sections to the best-positioned reader with a one-sentence rationale, e.g., "Your methods background positions you to assess the benchmark validity." Sections with no suitable participant are assigned to CoRead as an AI reader; A6 monitors and provides a grounded explanation (F3). Phase I can be skipped; CoRead then operates as a solo assistant (F4).

4.3 In-Session Orchestration (Phase II)

Phase II detects confusion and routes assistance without requiring readers to signal struggle publicly, addressing the social barrier all 12 formative participants described (F2).

AI Orchestration Core. A central orchestration core determines which agent to invoke, when, and in what sequence. On each signal update, the core evaluates D_s and selects exactly one action: suppress (fluent reading), notify via A7 (individual difficulty), route via A2 (peer-level), or activate group mode via A3 (distributed difficulty). This prevents simultaneous agent firing and enforces global cooldown windows. A6 is invoked only on acceptance; A8 monitors discussion independently of the routing pipeline (F2, F4).

Shared annotation layer (A4). All readers annotate the same PDF over a shared XFDF layer with colour-coded reason tags, inspired by Marshall's taxonomy [25]. Readers can hide others' annotations without affecting the shared layer (F4). A4 deduplicates highlights and surfaces divergence when two readers tag the same passage with conflicting reasons, turning interpretive disagreement into a coordination signal (F4).

Proactive assistance and Help Card (A7, A6). When D_s crosses τ_{fire}, A7 initiates three dismissible stages (F2), an amber margin indicator (Stage 1), a sidebar chip with a one-line hint 18 s later (Stage 2), and a full Help Card with diagnostic, element selector, and four-action grid (Stage 3). On acceptance, A6 generates

$$y = \text{LLM}(q \mid C_s) \tag{3}$$

offering three depths, *orientation*, *interrogation*, and *critical challenge* (P1, "explain like a beginner first"; P12, "I want a simpler idea first") (F2). Responses failing cosine-similarity grounding against C_s are withheld; peer routing is offered instead (F5). When multiple readers' D_s crosses τ_{fire} simultaneously, a shared confusion signal shows who shares the difficulty (F3). The *page pill* surfaces on single-reader D_s spikes, opening a location-aware diagnostic that classifies content type (notation, figure, table, prose) (F1, F2).

Silent verification, peer routing, and awareness (A8, A2, A3, A9). A8 monitors discussion for factual claims, attaching source badges (supported, contested, unverifiable) without interrupting (F5). When $D_s \geq \tau_{\text{high}}$ or A6 abstains, A2 routes to the peer with highest R_s using a live understanding score (score = $(1 - D_s) \times 100$, updated every 10 s), opening a chat where A3 seeds three paper-specific discussion prompts (F3); in the evaluation A3 prompts generated substantive discussion in 10 of 12 sessions. The sidebar supports inline `@AI`, group-level analysis, and active confusion markers (F3, F4). A9 surfaces citations when a reader annotates an unfamiliar term (F1). The Section Map combines struggle, highlight, and discussion events into a shared confusion heatmap (F3, F4).

4.4 Post-Session Stabilization (Phase III)

Phase III synthesises the full session record into a persistent wrap-up, addressing the post-session dimension of RQ3. Without it, clarifications disappear in chat and unresolved questions are forgotten. At session end, A5 generates a *Shared Understanding* document with four fields, *What we now know*, *Contested claims*, *Open questions*, and per-reader section summaries, each grounded in passage references (F3, F5). P3 and P8 rejected raw annotation logs as too noisy (P3 said "I don't want to see the whole process"; P8 reported that raw annotations "end up being more work"). Participants accessed the document a mean of 4.8 times per session, the highest of any feature. U5 observed, "In industry, the understanding just disappears after the meeting. This would solve that."

5 Evaluation

We conducted a single-condition feasibility evaluation addressing three questions. All quantitative results are reported descriptively to characterise feasibility and participant response, not to support population-level inference. (RQ1) whether the three-phase coordination produces the collaborative engagement patterns the formative study identified; (RQ2) whether behavioural signals support timely routing decisions and whether participants perceive those interventions as appropriate; and (RQ3) whether participants find the system usable. No baseline was included. This is a first deployment, and feasibility evaluations are conventionally scoped to establishing that mechanisms work before a comparative study [14]. REB File No. 19432; written informed consent and full anonymization.

5.1 Participants

Twelve participants (7 women, 5 men; age 22–41; U1–U12), distinct from the formative cohort, were recruited through purposive sampling from CS graduate programmes, 10 Master's students and 2 PhD students. The cohort provided the expertise heterogeneity needed to exercise peer-routing and role-assignment mechanisms. Participants were assigned papers counterbalanced by background. ML/NLP participants read Transformer-family papers (*Attention Is All You Need* [38], BERT [8], and GPT-3 [6]); those from adjacent fields read computer vision papers (CLIP [32] and ResNet [16]). Details appear in Supplementary Table A1.

5.2 Procedure

Twelve individual sessions were conducted over six consecutive days (March 25–30, 2026). Each session used a *Wizard-of-Oz protocol* [22] with a standardised confederate co-reader. This was a deliberate methodological choice. A feasibility evaluation of a multi-phase coordination system must first establish that the signal-to-routing pipeline is usable, acceptable, and interpretable before introducing the uncontrolled variance of simultaneous independent readers [20, 23, 30]. The study is thus scoped to validate the *mechanisms* (signal behaviour, notification acceptability, peer-routing function, usability), while the collaborative *dynamics* they support (concurrent routing, authentic social pressure, emergent group

study) are deferred to future work with unconstrained groups [14]. The standardised co-reader also held the surrounding social context constant, allowing grounding fallback and peer-routing behaviour to be examined without confounding variation from independent peer coordination.

The first author operated a second browser as the simulated co-reader under a minimal-activity protocol, three reason-tag annotations, two sidebar messages, and natural responses to system-initiated peer-routing. The confederate never initiated peer routing unprompted, so all peer-routing events reflect system-initiated routing triggered by the participant's D_s alone. Because the confederate was scripted and non-evaluative, real social pressure was absent; we return to this in the limitations.

After a 5-minute walkthrough, participants were told they could ignore any notification without penalty. WebGazer.js was calibrated with a 9-point procedure (target RMSE < 2.5°; achieved mean = 1.7°, SD = 0.28°). Participants then read their assigned paper for 44–67 minutes (mean = 52.3, SD = 6.4) under naturalistic conditions, engaging as they would for a reading-group discussion. After the session, each participant completed SUS [5], NASA-TLX [15], five 7-point Likert design-finding items (F1–F5), and a 10-minute semi-structured interview. Interviews were analyzed using codebook thematic analysis [3, 4], yielding four themes.

5.3 Measures

Collaboration quality (RQ1), peer exchange duration, Discussion Facilitator engagement (>5 min), Shared Understanding access count, peer-routing acceptance rate. *Struggle detection* (RQ2), intervention acceptance/dismissal, re-engagement after cooldown, grounding pass rate. *Usability and workload* (RQ3), SUS [5] (10 items, 0–100) and NASA-TLX [15] (0–100). *Design finding validation*, five 7-point Likert items per F1–F5. *Interaction logs*, all agent activations, routing decisions, accept/dismiss responses at 100 ms resolution.

5.4 Results

5.4.1 Collaboration Quality (RQ1). CoRead produced the collaborative engagement patterns the formative study identified across all three phases.

Phase I. All 12 completed their assigned sections and initiated at least one peer exchange within their domain. U4: *"The system had already told me what role they had, which made it easier to start."* U12: *"The role gave me legitimacy. Reading ML papers as a bioinformatician is normally isolating. This removed that entirely."*

Phase II. Of 13 peer-routing offers, 7 produced substantive exchanges. U2 had a 10-minute exchange explaining BERT; U8 described a GPT-3 scaling discussion as the most intellectually valuable part of the session. Discussion Facilitator prompts generated more than 5 min of substantive discussion in 10 of 12 sessions.

Phase III. The Shared Understanding sidebar was the most-used feature (mean = 4.8 accesses per session, SD = 1.3; all 12 used it at least three times). U3 noted, *"I kept checking what we had agreed on before moving to the next section. It changed how I read."*

5.4.2 Struggle Detection and Intervention Acceptance (RQ2). The Struggle Detector fired 53 AI notification interventions; 43 accepted (81.1%), 10 dismissed. Peer routing (triggered at τ_{high}) produced

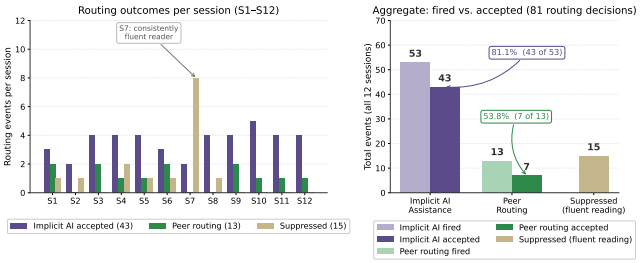


Figure 2: Routing outcomes per session (S1–S12) and aggregate fired vs. accepted. Implicit AI, 53 fired, 43 accepted (81.1%, F2). Peer routing, 13 fired, 7 substantive exchanges (53.8%, F3). Fifteen were suppressed during periods of fluent reading (F4). S7's high suppression count reflects one consistently fluent reader.

7 substantive exchanges from 13 events (53.8%). Including 15 interventions that were suppressed during fluent reading, 81 total routing decisions were recorded (Figure 2).

Two PhD participants (U3, U8) accepted a smaller share of interventions (3 of 6, 50.0%) than non-PhD participants (40 of 47, 85.1%); the sample is too small to support a population-level claim. U8 (PhD, NLP) explained, *"The threshold is clearly calibrated for a general audience, not for someone in this subfield,"* suggesting per-reader threshold adaptation is a direction for future deployment. Of 10 dismissed interventions, 6 were followed by re-engagement after the 90 s cooldown, confirming the persistence window filters transient from sustained confusion (F2). A6 was invoked 21 times; grounding checks passed on 19 of 21 (90.5%), both failures triggering peer routing via A2 (F5). These checks test lexical overlap, not human-judged fidelity.

Figure 3 shows D_s trajectories across three representative sessions. In S1, a novice crossed τ_{fire} at $t = 14$ min; AI explanation was accepted, A6 then abstained and routed to a peer at $t = 19$ min when grounding failed. In S3, the PhD reader never crossed τ_{fire}, confirming silence correctly suppresses intervention for expert readers. In S4, both readers crossed τ_{fire} simultaneously at $t = 8$ min, triggering the only group-study-mode activation across the 12 sessions.

5.4.3 Usability and Cognitive Load (RQ3). SUS ranged 68.5–87.5 (mean = 73.3, SD = 6.4, 95% CI [69.7, 76.9]), exceeding the 68-point threshold for all 12 [2]. NASA-TLX ranged 29.1–64.7 (mean = 47.0, SD = 10.9); mental demand dominated (mean = 55.0), reflecting paper difficulty rather than system-imposed load. Scores varied with domain familiarity (Figure 4).

5.4.4 Design Finding Validation (F1–F5). All five findings received mean ratings above the midpoint (Figure 5). F2 was lowest ($M = 4.5$): all 12 understood their role but 8 of 12 could not explain its basis. U1: *"Architecture Analyst made sense, but I didn't know why I got it. A one-line reason would make it feel less like a black box"* — a fix A10 already supports. F3 and F5 ($M = 5.7$) were joint-highest. F5 was driven by reduced social cost: U1 noted, *"I didn't have to admit out loud that I was confused; the system just understood."*

5.4.5 Feature Engagement and Qualitative Themes. Table 4 reports mean usage and satisfaction. Shared Understanding, inline equation

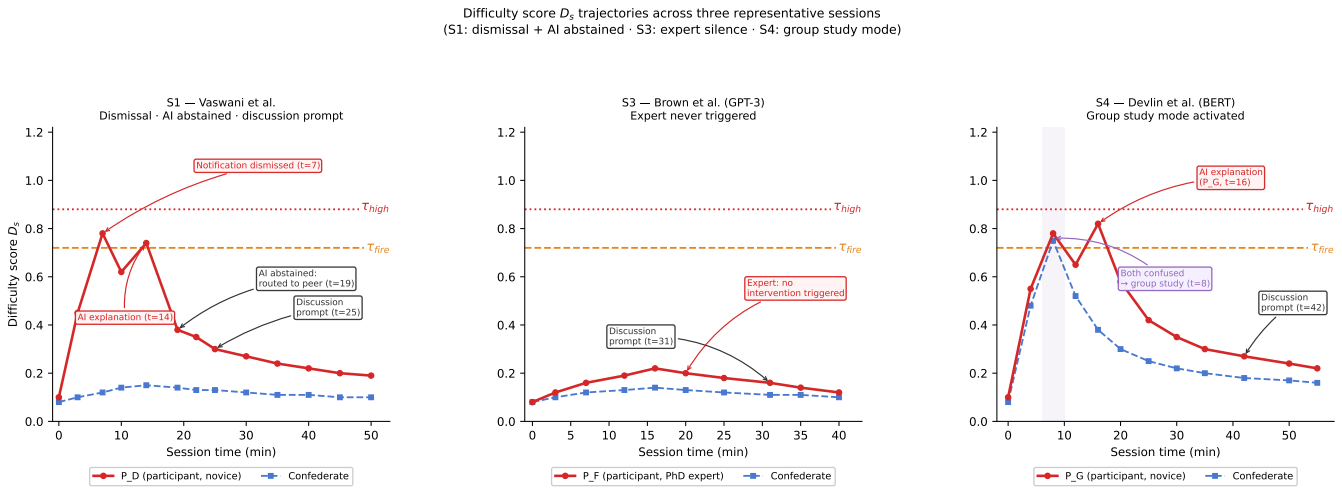


Figure 3: D_s trajectories across three sessions ($\tau_{\text{fire}} = 0.72$, orange dashed; $\tau_{\text{high}} = 0.88$, red dotted; participant solid, confederate dashed). S1, novice crossing τ_{fire} with notification dismissed at $t = 7$ min, then AI explanation at $t = 14$ min; A6 abstained at $t = 19$ min, routing to peer. S3, expert silence, PhD participant never crosses τ_{fire}. S4, both crossing simultaneously at $t = 8$ min, triggering the only group-study-mode activation observed. The trajectories illustrate how D_s changed around interpretable reading events during the feasibility evaluation.

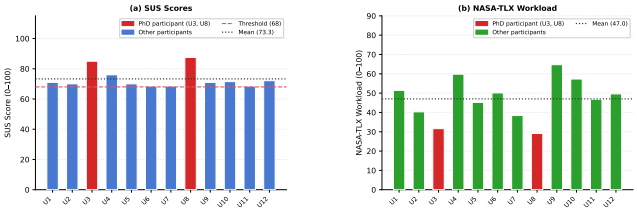


Figure 4: SUS (left) and NASA-TLX (right) per participant. PhD participants (U3, U8; red) had higher SUS scores (mean 86.2 vs. 70.7) and lower workload scores (mean 30.4 vs. 50.4) in this sample. Dashed line, 68-point threshold [2]; dotted line, group mean (SUS 73.3; NASA-TLX 47.0).

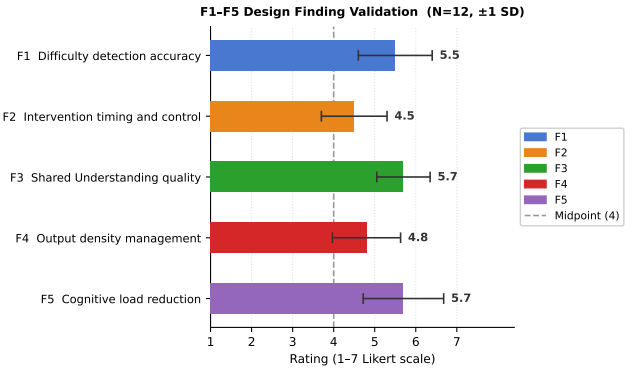


Figure 5: F1–F5 validation ratings ($N = 12$; 1–7 Likert). Error bars, ±1 SD. F3 and F5 joint highest ($M = 5.7$); F2 lowest ($M = 4.5$). Participants accepted intervention timing but expressed uncertainty about when the system chose to intervene.

explainer, and proactive sidebar formed the top-satisfaction cluster. Manual AI chat received the lowest satisfaction (3.4/5) despite high usage; 9 of 12 described responses as too long or poorly calibrated to their expertise, indicating an explanation register gap. Related Work Analyzer was discovered by only 2 of 12 (both PhD); A6 and peer routing covered the same terminological difficulties for the other 10 (F1).

Thematic analysis produced four themes. *Accurate at scale, imprecise at expertise.* Detection was accurate for non-experts but over-triggered for PhD readers. U9 said, *"It knew when I was stuck before I knew I needed help." The proactive sidebar as silent advocate.* Participants reported that implicit assistance removed the need to publicly signal confusion. U7 said, *"It felt like having a reading partner who notices when you slow down, without making it awkward." Shared Understanding as coordination artifact.* Participants described the sidebar as shared ground truth consulted mid-session. U11 noted, *"We kept coming back to it to check what we had settled. It was the anchor for the whole conversation." Role assignment, purpose without transparency.* All 12 reported their role gave a clear purpose, but 8 of 12 could not explain the basis for their assignment (F3).

6 Discussion

6.1 When and How: Answering the Two Design Questions

For when, stability-controlled routing produced 81.1% acceptance overall and suppressed intervention in 15 of 81 routing decisions during fluent reading (F2). These outcomes suggest that sustained behavioural evidence can support appropriately timed routing in a feasibility setting. The representative D_s trajectories in Figure 3 illustrate how the routing signal changed around interpretable reading events, but they should not be interpreted as evidence of classifier-level sensing accuracy.

Table 4: Feature engagement across all 12 sessions. Usage = mean accesses per session; Rating = mean satisfaction (1–5).

Feature	Usage	Rating
Shared Understanding sidebar	4.8	4.5
Manual chat with AI	4.1	3.4
Inline math/eq. explainer	3.1	4.6
Proactive sidebar	2.8	4.2
Figure/image explainer	1.8	4.5
Fact-check (auto)	1.2	4.2
Role card and assignment	1.0	3.8
Peer match notification	0.8	4.1
Research storyboard	0.5	3.5
Related work suggestions	0.2	4.5

For *how*, participants valued CoRead not primarily as an AI that answers questions, but as infrastructure deciding who should help whom. In single-user systems, the central question is when to interrupt the user [18, 21]. In collaborative reading, it becomes when to intervene in the group's shared coordination state, a different problem. AI in collaborative intellectual work should therefore be evaluated not only as an explanation tool but as coordination support, sensing when shared understanding is breaking down, routing help to the right person, and staying silent when the group does not need it [14, 39].

6.2 Participant-Reported Social Exposure During Help-Seeking

A recurring theme was that participants experienced CoRead's implicit assistance as less socially exposing than asking for help aloud in prior group-reading experiences. U1 said, *"I didn't have to admit out loud that I was confused."* U9 added, *"There is no one watching you ask a basic question. That matters more than people admit."* This non-public assistance mechanism is qualitatively distinct from annotation-based tools such as Perusall and NB, where confusion is visible by design [41]. CoRead provides both private handling through the notification path and chosen visibility through the shared annotation layer. The two serve different social functions and must coexist.

Caveat. This was observed under Wizard-of-Oz conditions; whether it holds under real social pressure remains open.

6.3 Design Guidelines

Five guidelines emerge for timely AI assistants in collaborative sensemaking, applicable beyond academic reading to any collaborative intellectual task with unevenly distributed expertise [14].

G1. Gate on sustained patterns, not isolated events. Assistance should require difficulty to persist across a temporal window before routing (F2).

G2. Preserve autonomy through non-modal, deferrable assistance. All system-initiated outputs must be dismissible at every stage without penalty (F2). Implicit monitoring must be opt-in.

G3. Route through peers before escalating to groups. Peer explanation is more trusted than AI for complex technical content; peer routing is less disruptive than group escalation (F3).

G4. Stabilise coordination into revisitable artifacts. Ephemeral chat does not constitute knowledge consolidation (F3, F5). Synthesise session activity into persistent, source-grounded artifacts.

G5. Ground all AI output in source material. Trust requires traceability (F5). Constrain generation to section-specific evidence and abstain rather than hallucinate.

6.4 Limitations

Three limitations bound this work. First, all sessions used a Wizard-of-Oz protocol validating mechanisms but not emergent group dynamics or real social pressure. Second, participants and papers came from CS; generalisation to other disciplines and non-expert co-readers is untested. Third, the evaluation reports downstream acceptance rather than direct sensing accuracy; precision-recall analysis against labelled confusion events and per-signal ablation are the primary next steps.

6.5 Future Work

Three directions follow directly from the limitations. First, deploying CoRead with real simultaneous reader groups will test concurrent routing, authentic social pressure, and group-study-mode activation under naturalistic conditions; a controlled comparison against a non-proactive annotation tool, a reactive AI assistant, and a no-routing CoRead variant would help isolate the contribution of the sensing-and-routing architecture. Second, extending the evaluation beyond CS would establish whether D_s weights and coordination structure hold across disciplines with different reading practices. Third, stronger grounding beyond cosine similarity would reduce A6 abstentions; direct sensing-layer validation via precision-recall analysis against labelled confusion events and per-signal ablation studies are immediate next steps.

7 Safe and Responsible Innovation

All signals are computed locally for D_s only; raw streams are never stored or transmitted and session data are deleted at session end. Participation is opt-in: Phase I can be skipped to disable tracking, readers can switch to solo mode at any time, and WebGazer is optional with graceful fallback to five signals. D_s thresholds are population-level estimates that may underperform for expert readers and underrepresented groups; per-reader calibration is a deployment priority. Use in evaluative settings such as grading is explicitly outside scope.

8 Conclusion

CoRead treats AI support for collaborative reading as a coordination problem: sensing when shared understanding breaks down, routing help to the best-positioned person, and staying silent when intervention would disrupt more than help. A Wizard-of-Oz evaluation ($N = 12$) showed 81.1% implicit intervention acceptance, 53.8% peer-routing acceptance, SUS = 73.3 (95% CI [69.7, 76.9]), NASA-TLX = 47.0, and grounding checks passing 19 of 21 explanations. These findings suggest AI in collaborative intellectual work should be judged not only by explanation quality but by coordination capacity: sensing breakdowns, routing support, and letting readers seek help without publicly announcing confusion.

Acknowledgments

The authors thank the study participants and reviewers. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant. LLM tools assisted with drafting; all technical content and claims were verified by the authors per ACM policy.

References

[1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM. doi:10.1145/3290605.3300233

[2] Aaron Bangor, Philip T. Kortum, and James T. Miller. 2008. An Empirical Evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[3] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[4] Virginia Braun and Victoria Clarke. 2021. One Size Fits All? What Counts as Quality Practice in (Reflexive) Thematic Analysis? *Qualitative Research in Psychology* 18, 3 (2021), 328–352. doi:10.1080/14780887.2020.1769238

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

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[7] Nancy J. Cooke, Jamie C. Gorman, Christopher W. Myers, and Jasmine L. Duran. 2013. Interactive Team Cognition. *Cognitive Science* 37, 2 (2013), 255–285. doi:10.1111/cogs.12009

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[9] Sidney D'Mello and Arthur 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

[10] Paul Dourish and Victoria Bellotti. 1992. Awareness and Coordination in Shared Workspaces. In *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work (CSCW '92)*. ACM, New York, NY, USA, 107–114. doi:10.1145/143457.143468

[11] Christopher Flathmann, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. 2021. Modeling and Guiding the Creation of Ethical Human-AI Teams. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM, New York, NY, USA, 455–465. doi:10.1145/3461702.3462573

[12] James Fogarty, Scott E. Hudson, Christopher G. Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C. Lee, and Jie Yang. 2005. Predicting Human Interruptibility with Sensors. *ACM Transactions on Computer-Human Interaction* 12, 1 (2005), 119–146. doi:10.1145/1057237.1057243

[13] Dongyang Guo, Yasmeen Abdrabou, Enkeleda Thaqi, and Enkelejda Kasneci. 2025. Multimodal Behavioral Patterns Analysis with Eye-Tracking and LLM-Based Reasoning. In *Proceedings of the 27th ACM International Conference on Multimodal Interaction (ICMI '25)*. ACM. doi:10.1145/3716553.3750787

[14] Carl Gutwin and Saul Greenberg. 2002. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work* 11, 3–4 (2002), 411–446. doi:10.1023/A:1021271517844

[15] 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.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778. doi:10.1109/CVPR.2016.90

[17] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S. Weld, and Marti A. Hearst. 2021. Augmenting Scientific Papers with Just-in-Time, Position-Sensitive Definitions of Terms and Symbols. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. doi:10.1145/3411764.3445648

[18] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM. doi:10.1145/302979.303030

[19] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of Attention in Computing and Communication: From Principles to Applications. *Commun. ACM* 46, 3 (2003), 52–59. doi:10.1145/636772.636798

[20] Kori Inkpen, Regan Mandryk, Joan DiMicco, and Stacey Scott. 2004. Methodologies for Evaluating Collaboration in Co-located Environments. In *Workshop at the ACM Conference on Computer-Supported Cooperative Work*.

[21] Shamsi T. Iqbal and Brian P. Bailey. 2008. Effects of Intelligent Notification Management on Users and Their Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 93–102. doi:10.1145/1357054.1357070

[22] John F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems* 2, 1 (1984), 26–41. doi:10.1145/357417.357420

[23] Keith Leavitt, Fangfang Qiu, and Debra L. Shapiro. 2021. Using Electronic Confederates for Experimental Research in Organizational Science. *Organizational Research Methods* 24, 4 (2021), 701–732. doi:10.1177/1094428119889136

[24] Yoon Lee, Haoyu Chen, Guoying Zhao, and Marcus Specht. 2022. WEDAR: Webcam-based Attention Analysis via Attention Regulator Behavior Recognition with a Novel E-reading Dataset. In *Proceedings of the 24th ACM International Conference on Multimodal Interaction (ICMI '22)*. ACM. doi:10.1145/3536221.3556619

[25] Catherine C. Marshall. 1998. Toward an Ecology of Hypertext Annotation. In *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia (Hypertext '98)*. ACM, New York, NY, USA, 40–49. doi:10.1145/276627.276632

[26] Catherine C. Marshall and A. J. Brush. 2004. Exploring the Relationship Between Personal and Public Annotations. In *Proceedings of the Joint Conference on Digital Libraries*.

[27] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Research. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[28] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/. Accessed: 2026-04-20.

[29] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI '16)*. AAAI Press, 3839–3845.

[30] David Pinelle, Carl Gutwin, and Saul Greenberg. 2003. Task Analysis for Groupware Usability Evaluation: Modeling Shared-Workspace Tasks with the Mechanics of Collaboration. *ACM Transactions on Computer-Human Interaction* 10, 4 (2003), 281–311. doi:10.1145/966930.966932

[31] Peter Pirolli and Stuart K. Card. 2005. The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. In *Proceedings of the International Conference on Intelligence Analysis*. McLean, VA, USA.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[33] Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin* 124, 3 (1998), 372–422. doi:10.1037/0033-2909.124.3.372

[34] Kjeld Schmidt and Carla Simone. 1996. Coordination Mechanisms: Towards a Conceptual Foundation of CSCW Systems Design. *Computer Supported Cooperative Work* 5, 2–3 (1996), 155–200. doi:10.1007/BF00133655

[35] Maia Stiber, Dan Bohus, and Sean Andrist. 2024. "Uh, This One?": Leveraging Behavioral Signals for Detecting Confusion during Physical Tasks. In *Proceedings of the 26th ACM International Conference on Multimodal Interaction (ICMI '24)*. ACM, 194–203. doi:10.1145/3678957.3685727

[36] Craig Tashman and W. Keith Edwards. 2011. LiquidText: A Flexible, Multitouch Environment to Support Active Reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, 3285–3294. doi:10.1145/1978942.1979430

[37] Carol Tenopir, Donald W. King, Sheri Edwards, and Lei Wu. 2009. Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. *Aslib Proceedings* 61, 1 (2009), 5–32. doi:10.1108/00012530910932267

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc., 5998–6008. https://papers.neurips.cc/paper/7181-attention-is-all-you-need

[39] Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes.* Harvard University Press, Cambridge, MA.
[40] Rui Zhang, Nathan J. McNeese, Guo Freeman, and Geoff Musick. 2021. "An Ideal Human": Expectations of AI Teammates in Human-AI Teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3, Article 246 (Jan. 2021), 25 pages. doi:10.1145/3432945
[41] Sacha Zyto, David Karger, Mark Ackerman, and Sanjoy Mahajan. 2012. Successful Classroom Deployment of a Social Document Annotation System. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems (CHI '12).* ACM, New York, NY, USA, 1543–1552. doi:10.1145/2207676.2208326