Attention is All We Need - Attention Recovery in E-Learning Using Webcam-Based Nudging

Gergely Horvath[1]

Vrije Universiteit Amsterdam

Abstract. Sustaining attention, particularly during complex reading tasks - requiring prolonged cognitive effort - is a challenge in the age of attention-diverting devices and distracting technologies, especially in e-learning environments relying specifically on said devices and technologies. This study investigates a real-time, webcam-based intervention system designed to detect attentional disengagement and deliver corrective nudges to refocus learners. The system developed uses computer vision to extract multi-modal behavioral signals: blinking, head orientation, and reading (eye movement periodicity), and based on these proxies it infers lapses of attention. Upon detection, the system "nudges" the user to redirect attention in accordance with Posner's attention reorienting model, either via a personalized goal-based text cue (top-down) or an impersonal humorous visual stimuli (an internet meme, bottom-up). The intervention aims to trigger cognitive re-engagement by leveraging mechanisms linked to sustained (tonic) and stimulus-driven (phasic) dopamine activity. The system effectiveness is evaluated with a between-subjects experimental design (N = 10), and its effectiveness is measured through reading comprehension and task engagement metrics. A subjective user questionnaire is administered before and after the session to collect qualitative data. This research contributes to the intersection of machine learning, cognitive neuroscience and educational technology, aimed at offering a playful, evidence-based scientific approach to adaptive attention support in digital learning environments.

Keywords: mind-wandering, educational technology, machine learning for attention detection, face detection, facial landmarks, e-learning, Posner's attention model

1 Introduction

The rapid shift to online education has underscored the importance of attention as a finite but critical cognitive resource for learning. In digital learning environments – where students face endless streams of multimedia content, notifications, and multitasking opportunities – sustaining focused attention is especially challenging. Cognitive scientists warn that unchecked Internet use and multitasking may erode attention, with some commentators claiming that ubiquitous digital access has a persistent detrimental impact on the brain's ability to concentrate. In Massive Open Online Courses (MOOCs) and other remote learning settings, this problem is acute: students often lack self-regulation to remain on task, contributing to the notoriously low completion rates (often around 10%) in such courses according to Onah et al. [21]. In fact, loss of learning focus (e.g., looking away from lecture videos or abandoning quizzes) is a core challenge in online education as it can significantly reduce learning efficiency. "Cognitive overload leads to confusion, distraction and disengagement." - according to the Digital Learning Institute [8]

Mind wandering and attentional lapses are empirically common in learning. Field studies report that students' minds wander 30–50% of class time [30,25], with lapses increasing under boredom,

fatigue, or stress — all of which hurt retention and comprehension. These fluctuations in attention — especially as a class or online lesson drags on — have obvious practical consequences: distracted students learn less. The societal stakes are high, given that large numbers of learners worldwide now depend on e-learning (a trend accelerated by the COVID-19 pandemic) to acquire skills and credentials. Addressing inattention is therefore not a minor detail but a central concern for improving educational outcomes in the digital age.

One promising approach is to build real-time adaptive systems that monitor attention via students' webcams and gently "nudge" them back on task. Modern computer vision enables off-the-shelf webcams to track gaze and head pose well enough to estimate where and how steadily a student is looking at the screen. Recent advances even make this approach scalable: for example, WebGazer [33], a JavaScript-based webcam eye tracker, can achieve sufficient accuracy and precision to predict students' task-unrelated thoughts (TUT) and comprehension in an online reading task. Similarly, Robal et al. [26] demonstrated that simply detecting whether a learner's face is visible in the webcam feed can serve as a coarse proxy for attention. In principle, such webcam-based cues could feed adaptive mechanisms: when the system infers that a learner's attention has drifted, it might trigger a nudge (e.g. a brief quiz or a pop-up reminder) to re-engage the student. Even basic implementations of this idea have shown promising results. In one experiment a Chinese high school used cameras to continuously score students' facial expressions for attentiveness; if a student's "attention score" fell below a threshold, the system notified the teacher, which reportedly led students to voluntarily improve their focus, as reported by Waldorf Today [32]. Real-world uses such as this one illustrate that continuous attention tracking can indeed elicit corrective behavior, for better or worse.

Designing effective webcam-based nudging systems calls for integrating insights from cognitive psychology. Classic models of attention (e.g. Broadbent's early filter theory [4]) underscore that human information processing has strict capacity limits: unattended inputs may be filtered out entirely (also supported by Lachter et al. [17]). More recent frameworks (e.g. Posner's attention networks [23]; Engle's executive-attention models [10]) emphasize that attention is regulated by multiple neural systems responsible for alerting, orienting, and top-down control. In the learning context, these systems must constantly battle against mind-wandering and distraction. Importantly, neuromodulatory signals like dopamine play a key role in this battle. Dopaminergic pathways encode reward and salience to motivate behavior: phasic dopamine-bursts in midbrain neurons signal reward prediction errors and reinforce attended stimuli, while tonic dopamine establishes baseline motivational vigor [5,29]. In other words, a student's sustained attention depends not only on moment-to-moment stimulus properties but also on the larger motivational context signaled by dopamine. An adaptive learning system that detects waning attention could, in principle, compensate by introducing motivational "rewards" (for example, positive feedback or gamified elements) to boost the student's engagement.

This study builds on these considerations to design and test a webcam-based nudging system for attention recovery in e-learning. By continuously estimating learner engagement with a standard laptop camera and delivering timely nudges (prompts, content adjustments, or motivational messages), such a system could help students maintain focus. The feasibility of this vision is supported by initial studies in both sensing and intervention. However, despite growing interest, research on webcam-driven attention nudging in e-learning remains fragmented. The following sections will review relevant literature in detail, covering cognitive models of attention and motivation, the prevalence and impact of mind-wandering on learning, and technical advances in webcam-based attention sensing, affective education, and language-model-driven adaptivity.

2 Related Work

Cognitive Models of Attention Theories of attention have evolved over decades, from early selection filters to multi-component control systems. Broadbent's classic filter model (1958) [4] posited that sensory information is initially processed only by basic physical features, with a selective "filter" allowing one channel to reach higher processing; unattended stimuli receive no semantic processing. This idea was later challenged by findings of "leakage" of unattended information, but Lachter et al. [17] showed that under strictly controlled conditions, truly unattended stimuli do not prime cognitive processing – effectively supporting a revised version of Broadbent's filter. Successor theories introduced flexibility: Treisman's attenuation model [31] allowed some reduced processing of unattended streams, and load theory suggested that high perceptual load in the attended task forces stronger filtering. Posner and Petersen [23] framed attention in terms of separable brain networks – an alerting system that maintains a vigilant state, an orienting system that shifts attention spatially or to features, and an executive or control network for resolving conflict and maintaining goals. The executive attention network—sometimes equated with working-memory capacity—is especially relevant for resisting distractions and recovering focus when mind-wandering intrudes. Together, these classical and modern models converge on the view that attention involves both early selection and late control stages, with limited capacity and reliance on top-down regulation.

Dopamine, Motivation, and Cognitive Control. Neuroscience has linked dopamine (DA) to both reward processing and attentional engagement. Midbrain DA neurons exhibit two firing modes: tonic firing maintains a baseline level of dopamine that tunes neural circuit excitability, while phasic bursts (100—500 ms spikes) produce transient dopamine surges in response to salient events [5]. Phasic DA signals are triggered by unexpected rewards or reward-predictive cues and serve as a teaching signal in reinforcement learning. In educational terms, a correctly solved problem or an encouraging feedback might induce a phasic dopamine response that reinforces focus on that task. Theoretical accounts have expanded this view: Bromberg-Martin et al. note that distinct DA neuron subtypes encode value versus salience signals. Value-coding neurons increase activity for rewards and reduce it for aversive events, supporting goal-directed behavior. In contrast, salience-coding neurons respond to both positive and negative events, supporting arousal and vigilance. Both pathways, augmented by a brief DA "alerting" signal to any unexpected cue, work together to drive adaptive behavior. Thus, DA plays a critical role in linking motivation, attention, and executive control. When students find content rewarding or salient, dopamine pathways help sustain focus; conversely, when engagement wanes, a lack of phasic reward signals can lead to disengagement. Modern attention-control theories often incorporate dopaminergic concepts (e.g. cognitive effort models), highlighting the interplay of attention and motivation in learning.

Mind-Wandering in Learning. Mind wandering - often defined as task-unrelated thoughts (TUTs) - is a major form of attention lapse. Pioneering surveys and lab studies show that people spend a large fraction of waking life mind-wandering, with estimates of around 30–50% depending on context. In educational settings, Kane et al. [15] review evidence that mind-wandering is very common in lectures and online courses. They report that on average students' minds wander about 30% of the time, increasing under fatigue or boredom [30]. Strikingly, Risko et al. [25] find that mind-wandering increased from 30% in the first half of a lecture to 49% in the second half, and that episodes of wandering in the latter half were associated with significantly worse test performance [30]. These results echo other findings (e.g. risk of mind-wandering in lengthy videos, and

daily-life experience-sampling studies) showing that most learners zone out periodically unless actively engaged. The negative impact of mind-wandering on learning is well documented: periods of inattention lead to missed information, lower comprehension, and poorer encoding of material. This has prompted research into interventions to reduce TUTs in class (such as interspersed quizzes or prompts). All in all, mind-wandering theory and empirical data confirm that sustaining attention is not automatic – learners' attention frequently drifts, and those drifts measurably impair learning outcomes.

Webcam-Based Attention Estimation. Recent human-computer interaction (HCI) and educational technology work has explored the use of webcams to infer user attention. Before deep learning, most webcam attention systems used handcrafted features and traditional ML classifiers. For example, Ross et al. [27] applied K-means clustering and Support Vector Machines to data from an RGB-D camera (tracking head and body posture) to classify students as "attentive" or "inattentive". In the web context, Robal et al. evaluated simple webcam analytics: they monitored whether a learner's face and gaze were on the video of a MOOC lecture, treating absence or diversion as a proxy for inattention [26]. Their results highlighted the difficulty: even off-the-shelf face trackers struggled to reliably detect gaze on small laptops, and simpler face-present heuristics often lagged by several seconds. These early attempts demonstrate proof-of-concept but also motivate more sophisticated solutions.

In recent years, deep learning has greatly improved webcam eye- and gaze-tracking. Convolutional neural network (CNN) models can be trained to map facial images to gaze coordinates without special calibration. For instance, Saxena et al. (2024) reported that deep learning gaze models achieved average gaze errors of $\sim 2.4°$, substantially narrowing the gap between webcam and lab eye trackers [28]. Similarly, models like *WebGazer* use regression-based gaze estimation calibrated on user data to infer screen-look with acceptable precision [33]. Beyond gaze, multi-modal neural models have been explored: SVD-based CNNs on facial video have been proposed to predict mind-wandering directly [1]. The trend is clear: where classic ML required specialized hardware to coarse features, deep networks enable more accurate, real-time attention estimation from everyday webcams. While screen-centered gaze offers a practical cue, it imperfectly represents attention since learners can visually fixate yet mentally disengage. Incorporating temporal models like LSTMs or Transformers could better capture attention dynamics by analyzing patterns over time rather than static snapshots. Either way, it remains challenging to translate raw gaze into cognitive state; most systems still rely on relatively simple heuristics (e.g. "eyes off screen" or blink rate) and await more contextual understanding.

Affective Computing and Adaptive Learning Systems. Detecting attention is closely related to detecting affective engagement [34]. Affective computing frameworks have long aimed to recognize user emotions (via facial expression, posture, vocal tone, or physiological signals) and adapt interfaces accordingly [14,19]. In education, researchers have built emotion-aware intelligent tutoring systems (ITS) and affective tutoring systems (ATS) that monitor learner engagement and frustration to guide interventions [2,16]. For example, if a student's face indicates boredom or confusion, the system might present a hint or a motivational message. Meta-analyses and reviews suggest that integrating affective cues can improve learning outcomes: Linnenbrink Garcia et al. note that positive valence and arousal (detectable through facial cues) tend to enhance cognitive flexibility and performance. In practice, educational platforms are increasingly incorporating simple facial analytics: some modern ITS adaptively select content based on user-reported or inferred

engagement, and a few experimental MOOCs have trialed live video analysis to detect off-task behavior. Although privacy and ethics remain concerns, the technical feasibility of real-time facial analysis (e.g. using standardized APIs for emotion recognition) is now established. These adaptive systems overlap with our focus on attention: attention and affect share underlying signals, so many affect-sensitive tutors can also function as "attention-aware" tutors.

Large Language Models and Adaptive Messaging. The recent rise of generative language models opens new possibilities for personalization. LLMs (e.g. GPT-3, ChatGPT, BERT variants) can process student input (questions or essay drafts) and generate custom responses, hints, or explanations. A 2024 review observes that LLMs have become ubiquitous as virtual tutors for tasks like question generation, answer evaluation, and automated feedback [11]. These models possess broad world knowledge and natural language understanding, enabling them to engage learners conversationally. For example, an LLM could interpret a student's chat question ("I'm confused about step 2") and produce a tailored clarification, or even adjust its tone and complexity based on the learner's level. Recent proposals for "LLM-based education systems" suggest two modes: a unified AI tutor or a mixture-of-experts ensemble, each driven by a language model core [18]. The key advantage is flexibility: unlike static rule-based systems, an LLM can potentially generate an almost unlimited variety of prompts or motivational messages. In the context of attention recovery, LLMs might be used to craft adaptive nudges. For instance, upon detecting low engagement, the system could invoke an LLM to produce personalized encouragement or to reframe the learning goal in the student's own words, thereby reigniting long-term goals and motivation. While still emerging, early work on LLMs in education shows promise (and challenges) in dynamically responding to student needs.

3 Research Question

"How effective are nudges (short-form meme videos vs. personalized goal-oriented text-based cues) in restoring attention after disengagement during a reading task?"

4 Methodology

4.1 Overview

This study developed and tested a webcam-based attention monitoring system for e-learning. Using real-time computer vision via MediaPipe [20], the system tracked behavioral cues (gaze, blinking, microsaccades) to estimate attention levels. When attention dropped below a certain threshold, the system delivered one of two types of nudging interventions depending on the group assignment:

- Goal-based textual cues (Group 1)
- Humorous meme videos (Group 2)
- No nudges (control) (Group 3)

All computation and data storage occurred locally. The objective was to determine whether such nudges could facilitate attentional recovery and improve comprehension in a controlled reading task, and to measure the effectiveness of such interventions between the three groups.

4.2 Nudging Details

Nudges shown as a pop-up on-screen text (Group 1) or a pop-up short meme video (Group 2). Bell sound played for Group 1 to match stimulus salience of Group 2's video nudges. Both nudges were reactive only (i.e., not predictive). A nudge remained visible until actively dismissed by the user, and only one nudge could appear on-screen at any given time to avoid cognitive overload. The number of nudges per session varied, but the system allowed for a limitless amount of them (in this case, as long as the messages were not exhausted from the list so as to avoid duplicate notifications).

4.3 Participants

- Random sample of 10 university students
- Exclusion: low English proficiency, extreme fatigue (<4h sleep)
- Randomized (non-stratified) group assignment

5 Experimental Design

System details are explained more precisely after this section.

5.1 Pre-Experiment

Participants completed a GDPR consent form and a pre-task questionnaire assessing the following dimensions:

- **Demographics:** age, sex, sleep duration
- **Experience:** prior exposure to *Moby Dick*, reading habits, English proficiency
- **Self-assessment:** distraction susceptibility, focus retention, goal orientation
- **Psychological traits:** procrastination, impulsivity, intrinsic motivation

Participants in Group 1 were also asked to collect and reflect on their personal goals. These responses were later used to generate individualized nudging messages.

5.2 Task Execution

Each participant launched the experiment program, which initialized the webcam feed and loaded the attention-monitoring model in the background. Subsequently, the first three chapters of *Moby Dick* were displayed in PDF format. Webcam-based signal processing began immediately. Attention levels were computed per frame using a heuristic model (see next section).

Intervention conditions:

- **Group 1**: Received goal-based nudges accompanied by a bell sound when attention dropped below 0.1.
- **Group 2**: Received a short meme video (with sound) as the nudge under the same attention threshold.
- **Group 3**: No intervention was triggered.

Each intervention remained active until the participant closed the corresponding pop-up window.

Participants were free to read for approximately 30 minutes. While reading duration and progress were logged, completion was not enforced. Prior to the session, participants were primed with a brief, calming conversation to minimize performance anxiety. Researcher left participant alone during task.

5.3 Post-Experiment

Upon completing the reading session, participants were presented with a short multiple-choice reading comprehension test consisting of approximately 15 questions (5 per chapter). Participants could select "I don't know" as an answer, which served a dual purpose: enabling honest response behavior and acting as a proxy for estimating reading progress, as they were instructed to select it when they had not yet reached a given section.

Following the test, a brief qualitative questionnaire captured subjective participant feedback. It included items on perceived engagement, instances of mind-wandering, perceived false positives in attention detection, and general impressions of the system. The final item was an open-ended text box allowing for unrestricted commentary on the experience.

6 Data Collection

6.1 Quantitative Data

During the experiment, the attention system recorded the following metrics on a per-frame basis:

- `is_reading`: Discrete value ranging from –1 to +1 (proxy for engagement with text).
- `is_looking`: Discrete value ranging from –1 to +1 (indicates gaze alignment with screen).
- `is_blinking`: Binary indicator (0 or 1) of eye closure.
- `attention_level`: Instantaneous frame-level attention estimate.
- `attention_score`: Rolling average of recent `attention_level` values.

Following the reading task, participants completed a multiple-choice comprehension quiz. Their responses were stored and used to evaluate reading success. Each quiz item corresponded to specific sections of the text, allowing correlation with attention trends over time.

6.2 Qualitative Data

In addition to quantitative metrics, participants completed a post-task qualitative questionnaire assessing subjective experience across the following dimensions:

- **Engagement level**: How immersed or interested the participant felt during reading.
- **Perceived difficulty**: Subjective evaluation of text complexity.
- **Focus level**: Self-assessed ability to maintain concentration.
- **Mind-wandering**: Frequency and severity of zoning out, as reported by the participant.
- **Intervention helpfulness**: Perceived utility of the nudging/meme-based interventions (Groups 1–2).
- **Personal relevance of nudges**: Whether the messages felt tailored and motivating (Group 1 only).
- **Willingness to reuse system**: Openness to engaging with similar systems in future reading tasks.
- **Open comments**: Free-text section for additional feedback, remarks, or complaints.

6.3 Data Analysis Plan

To assess the efficacy of the attention-intervention system, the data collected during experimental test sessions involving three participant groups are examined:

- **[T] T**ext-based, goal-oriented, nudging
- **[M] M**emes: goal-irrelevant, humorous (video) nudging
- **[C] C**ontrol group without intervention

The system logged the following variables:

- `time` (continuous float)
- `attention_score` (float)
- `is_blinking`, `is_looking`, `is_reading` (boolean)
- `nudge` (boolean; indicates intervention occurred)

The intervention is considered successful if the nudging groups [T] and [M] outperform the control group [C] across the following dimensions:

- Higher percentage of time spent in the attentive state
- Shorter average duration of inattentive episodes
- Longer average attentive periods following a nudge
- Greater reading progress within fixed time intervals
- Higher comprehension scores on post-reading tasks
- More positive subjective feedback regarding experience

Attentive State Percentage. The mean proportion of time spent in the attentive state is compared across groups. A significantly higher percentage in [T] and [M] compared to [C] would indicate improved sustained attention.

Inattentive State Duration. Analyzing the average length of inattentive episodes. Lower durations in the intervention groups would suggest that nudges successfully re-engage attention more rapidly.

Nominal Attentive Timewindows. For each detected attention drop, the average duration of the subsequent attentive state is measured. In the control group, simulated "would-be" nudge points are placed for comparison. Longer attentive periods following actual nudges ([T] and [M]) indicate successful re-engagement.

Reading Progress and Comprehension. The amount of material read in a fixed period and the accuracy on related comprehension tasks is quantified. Superior scores in the intervention groups support the effectiveness of nudging in facilitating learning.

Subjective Feedback. Post-task questionnaires assess user experience. Differences in perceived helpfulness, clarity, or intrusiveness are evaluated across groups using Likert-scale analysis.

7 System Details [12]

Key design principle: This system distinguishes between instantaneous perceptual inference — termed *attention level* — and its temporally aggregated counterpart, *attention score*.

- **Attention level** is a discrete, ternary variable $\in \{-1, 0, +1\}$ computed per frame based on heuristics applied to three primary features: head position, blink detection, and reading activity.
- **Attention score** aggregates attention levels over a rolling temporal window, producing a smoother, more stable estimate of learner engagement.

7.1 How attention is measured - Overview

In order to detect and infer lapses in learner attention during e-learning sessions, the system utilizes a webcam-based multi-feature monitoring system. The approach draws on empirical literature discussed previously identifying blinking, periodic pupil movements, and head orientation/gaze direction as reliable, noninvasive background indicators and proxies of cognitive engagement. Each of these features are independently extracted in real time from webcam footage using computer vision pipelines based on existing open-source frameworks (in this case, Google's MediaPipe [20]), then processed to identify deviations from an engaged cognitive state.

7.2 Feature Selection and Rationale

Head Orientation and Gaze: Head orientation and gaze direction serve as nonverbal cues of learner focus, Consistent with findings from Buono et al 2023 [6] and Li & Liu 2024 [18]. Attention is inferred to be high when the user's head is upright and gaze is centered towards the screen. Conversely, frequent lateral head rotations, downward tilts, or off-screen gazes are flagged as potential distractions.

Blink Behavior: Eye-blink behaviour has been widely linked to cognitive load and attention state. Several studies (e.g.: Riby et al. 2024 [24], Paprocki & Lenskiy 2017 [22]) demonstrate that elevated blink rates correlate with mind-wandering episodes, whereas lower and more stable blink frequencies tend to co-occur with sustained task engagement. To take advantage of this, the system employs a simple heuristic: The more a user blinks, the lower the attention.

Reading Activity: The most concrete and directly relevant feature however is the intricate ways the eyes move during the reading task. Regardless of blink frequency or head motion, without active reading, no textual information is assimilated. Reading tasks are characterized by rapid, rhythmic eye movements (saccades and fixations) as the reader progresses through textual material. In order to detect such patterns, the pupil points as well as the eye points were selected.

All selected features 1 were calculated to be agnostic of the absolute position of the user with regards to the screen, and only the relative distances were calculated. Thiss also anonymizes data, eliminating any personally identifiable positional information from the collected set.

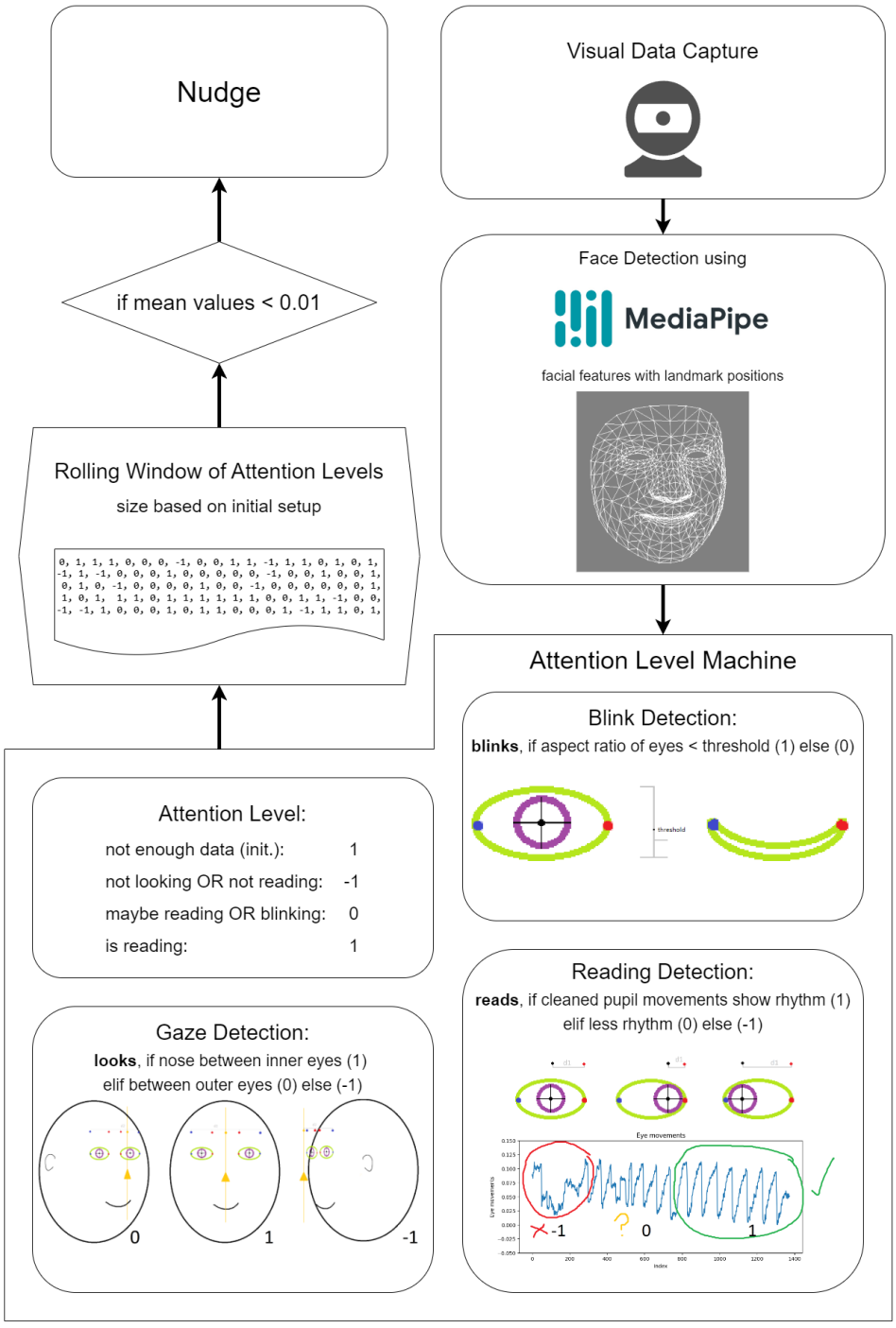

Fig. 1: System Design

7.3 Visual Signal Extraction and Processing

Head Pose and Gaze Lateral head orientation is estimated using landmarks on eye corners and the nose tip. Conceptually: with the nose between the inner eye points as certainly looking at the screen, nose between outer eye points as unsure and outside as certainly not looking at the screen. Formally, we define:

$$\text{is\_looking} \in \{-1, 0, +1\}$$

where $+1$ denotes engaged gaze toward the screen, 0 ambiguous, and -1 disengaged gaze.

Blink Detection Blink rate correlates with attention: higher rates suggest mind-wandering. The system computes the eye aspect ratio (EAR), and sudden EAR drops (blinks) reduce the attention estimate. Heuristics guard against false positives due to lighting or pose changes (in case the eyes become obstructed or ambiguous).

Formally:

$$\text{is\_blinking} \in \{0, +1\}$$

where $+1$ indicates a blink event detected at the frame.

Pupil Tracking Pupil tracking is a necessary feature, as reading itself induces rhythmic horizontal pupil motion (saccades and fixations) while the reader progresses through textual material. The system computes normalized horizontal displacement of the right pupil using MediaPipe landmarks, measured as a ratio relative to eye width. This signal is stored in a rolling buffer and preprocessed with:

- 4th-order Butterworth low-pass filter to suppress high-frequency noise
- Linear detrending to remove baseline drift
- Application of a Hamming window followed by Fast Fourier Transform (FFT)

If the dominant frequency lies in the 0.2–3 Hz band with sufficient amplitude, reading is inferred. This value has been reached through rigorous empirical testing, but can be subject to improvements. The length of the rolling buffer sets a temporal leniency parameter, tunable based on empirical performance.

Each frame receives a discrete reading classification:

$$\text{is\_reading} \in \{-1, 0, +1\}$$

where $+1$ means certainly reading, 0 ambiguous, and -1 certainly not reading.

Heuristic Summary: The process can be informally restated as: pupil displacement ratios form a time series treated like a sound wave; after filtering and detrending, the FFT identifies the dominant frequency. If this frequency aligns with typical reading frequency (neither zero lines per hour nor 100 lines per second) and is prominent compared to noise, the user is classified as reading ($+1$). Ambiguous frequency or amplitude yields a 0, and other cases receive -1.

Classification thresholds are:

- Dominant frequency $\in [0.2\,\text{Hz}, 3.0\,\text{Hz}]$ and amplitude ≥ 0.5: **actively reading** ($+1$).
- Dominant frequency $\in [0, 4.0\,\text{Hz}]$ and amplitude ≥ 0.3: **ambiguous** (0).
- Otherwise: **not reading** (-1).

This lightweight, real-time approach requires only webcam input and minimal computation.

7.4 Frame-wise Attention Classification - Attention Level

For each frame all three of these metrics are calculated. Then a simple heuristic is used to determine (or more precisely, approximate) a discrete momentary user attentiveness. The heuristic is the following:

- **+1** — Engaged reading: consistent gaze direction, valid reading periodicity, and eyes open
- **0** — Partial engagement: ambiguous signals or blink events
- **−1** — Inattentiveness: no reading signature, off-screen gaze, or prolonged eye closure

Formally,

$$\text{attention\_level} \in \{-1, 0, +1\}$$

These values are stored sequentially in a rolling buffer for temporal aggregation.

7.5 Temporal Aggregation - Attention Score

This design separates immediate perceptual inference (attention level) from temporal state estimation (attention score), allowing responsive yet robust attention tracking.

Rolling Attention Score The temporal aspects of attention were accounted for based on pedagogical literature, suggesting that attention typically declines over time, most often within 10-15 minutes of sustained activity (Bradbury, Neil A., 2016 [3], Darnell, D.K. and Krieg, P.A., 2019 [7]). However, empirical findings vary, and abrupt declines after 1-2 minutes are not consistently observed. Therefore, rather than relying on fixed temporal assumptions, as well as due to the timely limitations of the experiments, this system adopts a fully data-driven approach, where engagement signals are evaluated continuously.

To stabilize noisy fluctuations in the real-time signal, a rolling attention score was calculated using a 15-second ($\tilde{4}50$-frame under 30FPS) moving average of the frame-wise attention level:

$$\text{attention\_score}(t) = \frac{1}{N} \sum_{i=0}^{N-1} \text{attention\_level}(t - i)$$

where N corresponds to the 15-second window length.

Nudge Trigger Logic This **attention score** served as the decision metric for intervention logic:

- When the rolling attention score (the average of the calculated attention levels for every single frame during the past 15 seconds) dropped below a predefined threshold (0.1), an attention-nudging mechanism was triggered.

This threshold also captures user states such as dozing off—where blinking events ($\text{is\_blinking} = 1 \implies \text{attention\_level} = 0$) accumulate, driving the rolling average down to zero and triggering the nudge.

7.6 Limitations and Excluded Channels

Facial Emotion (Excluded) A channel of engagement inference that was excluded is facial affect recognition. Building on previous work (e.g. Gupta et al. 2023 [13], Buono et al. 2023 [6]), a CNN could be used to classify facial expressions into discrete emotional categories (e.g. neutral, happy, sad, bored, surprised). Prior research indicates that negative affect - especially boredom, sadness and confusion - often precedes or co-occurs with attention lapses in learners (D'Mello, Sidney, et al. [9]). The model could have output a categorical emotion prediction at each frame, which then would have aggregated over time to estimate the user's emotional state. However during the development of the system, it became obvious through empirical experiments that there is no one-size-fits-all solution to the relationship between facial emotion and user attention. For one, the author of this paper tends to have a rather angry look on his face, and is often confused - at least according to available facial emotion detection models. Though there may be ways to integrate emotion scores as well, it was decided to exclude this aspect from the final result.

Temporal Decay Models (Excluded) Rather than hard-coding attention decay curves (e.g. 10–15 min drop-off), the system relies on continuous signal-driven estimation, whereby though avoiding assumptions about user behavior, at the same time assuming great performance on its own end.

8 Results

8.1 Observations

Interventions were noticed and generally provoked frustration. Some participants deliberately tried to avoid nudges, especially memes. Ironically, this avoidance behavior led to higher engagement and focus, as participants self-corrected to not trigger the interventions

8.2 Kruskal-Wallis

To compare average attention levels across three experimental groups, a Kruskal–Wallis H-test was used on each participant's mean attention score, which had been calculated from their full attention time series. Participants were grouped according to the last digit of their ID number ("1", "2", or "3"), resulting in group sizes of 3, 3, and 4. The test did not find a statistically significant difference between groups: $H(2) = 2.200$, $p = 0.333$. This suggests that average attention levels, when ranked, were not consistently different between the groups. However, the small sample sizes and noticeable variability—such as a negative z-scored mean in Group 2—indicate that results should be interpreted with caution. A boxplot (see Fig. 2) was used to visualize the group distributions. Though there are some visible differences in central tendency, these are not statistically significant. Future research with more participants and possibly bootstrapped confidence intervals is recommended to better evaluate patterns in attention levels.

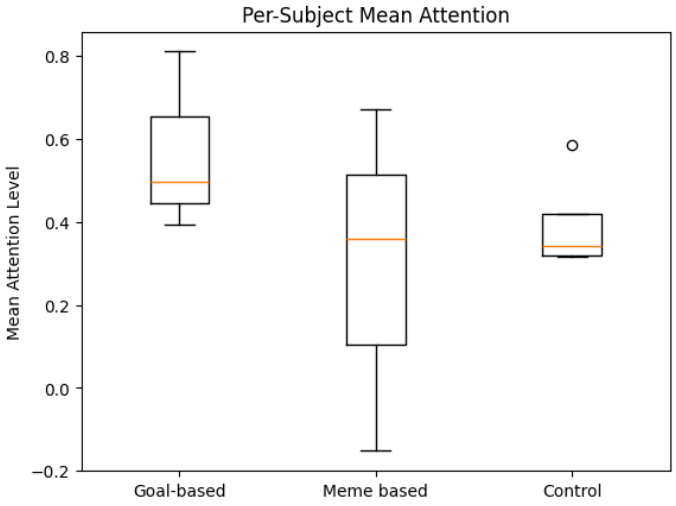

Fig. 2: Boxplot of per-subject mean attention scores by group.

8.3 Individual Results Ranked 10.2

As defined in subsection 6.3, *Attentive State Percentage*, *Inattentive State Duration*, *Nominal Attentive Timewindows*, *Reading Progress*, and *Reading Comprehension* are used as quantitative metrics of system performance. The corresponding boxplot visualizations are provided in Appendix 10.2.

The most notable trend in the data is that text-based, goal-oriented nudging outperformed both the meme-based and the control groups performance in all five metrics, with the outstanding differences measured in *Attentive and Inattentive Durations*.

Surprisingly, the meme-based intervention decreased the average *Attentive State Percentage* by 9.5%. This suggests that during reading, this form of nudging has introduced a form of distraction instead of reinforcing task focus - precisely the opposite of the intended effects of the system.

Metric	**Control**	**Meme**	**Text**	**Δ% (Meme)**	**Δ% Text)**	**Table**
Attentiveness % ↑	77.93	70.50	83.25	−9.5%	**+6.8%**	1g
Inattentive Duration ↓ (s)	22.61	6.74	6.43	**-70.2%**	**-71.6%**	2g
Attentive Duration ↑ (s)	46.60	101.91	164.21	**+118.8%**	**+252.4%**	3g
Reading Correctness % ↑	43.22	52.38	70.21	+21.2%	**+62.5%**	4g
Reading Length % ↑	31.95	20.37	48.15	-36.2%	**+50.7%**	5g

Table 1: Intervention Effects Relative to Control Group (↑ higher is better, ↓ lower is better)

8.4 Qualitative Observations 4, 5

System Sensitivity Both intervention groups noted that the system was sometimes too sensitive or gave false positives, nudging them despite focused attention. Glasses and ambient distractions (e.g., roommates) were mentioned as possible confounding factors explaining the oversensitivity.

Emotional Reactions to Nudging Almost all participants reported spite-based overcompensation towards the nudges; they were trying to beat the system by focusing more just to avoid nudges. The responses of the text group were more reflective, and the responses of the meme group leaned toward annoyance or disengagement.

Personalization & Context Awareness Several users expressed that static nudging logic does not work well without context (e.g., comprehension vs. gaze). Users suggested that the system would improve if it could "get to know them" or adapt to real-time comprehension issues. Text-based nudging seemed to have more personal connotations and was reported to be more helpful than Memes.

Humor Meme group was critical of the content quality:
"Videos were not funny"
"Took me longer to get back to reading after laughing"

Positive Feedback Exists but Is Nuanced At least one user found the tool potentially useful for exam prep, noting the value of urgency. Another appreciated the post-reading questionnaire more than the nudge system itself.

9 Discussion

Attention is a complex cognitive process, that is not necessarily a straight-forwardly measurable activity with simple webcam-based techniques. Early results suggest that gaze-, blink-, and reading-tracking can correlate with an improved reading task performance, however, despite objective quantitative improvements, user responses indicate strong feelings against the system and its usefulness, citing annoyance and spite as significant hindering forces to prolonged use. This alone indicates that while this system could improve productivity in mechanised (structured) environments like schools or workplaces, there are ethical concerns in institutionalizing attention-tracking systems against the will or beyond the comfort levels of users.

9.1 Methodological Limitations

There is a methodological limitation in the current setup of the experiments, in that only the text-based goal-oriented group was primed with questions about their life-goals and the ways literacy can help in achieving them. This introduced a confounding variable, as the act of goal reflection and articulation itself might enhance attention and comprehension by a measurable amount, independently of nudging or other intervention. Future studies should control for this priming effect, isolate the specific contribution to attention improvement of the nudges themselves.

Another confounding variable has to be mentioned: participants were aware of being monitored, which, although uniform across groups, most certainly has influenced behaviour (via a Hawthorne effect). The effect was partially mitigated by painting the webcam indicator light black, in order to avoid the constant awareness of the surveillance, but it is unclear how effective this mitigation turned out to be.

9.2 Relation to Attention Models

These early results coincide with Posner's model of attention systems, and suggest that goal-driven, top-down attentional control (in this case, personalized text nudges) have stronger influence on behavior than stimulus-driven, bottom-up signals (in this case, meme videos). Even with the small sample size (N=10), the main trend of text nudges outperforming meme nudges provides a baseline foundation for further explorations.

10 Future Work

10.1 Adaptivity and Personalization

Based both on initial assumptions and backed by user feedback, a most crucial frontier of future research must be more adaptivity and deeper personalization, that both relies on (for data collection) and simultaneously facilitates (by the use of collected data) long-term use. Implementing dynamic, user-tailored nudging content requires prolonged interaction and much feedback. But an initial phase of a certain duration, where the system would track attention patterns (e.g., average duration before distracion) and interactively refine nudging timing and content, perhaps even predictively rather than reactively, would most definitely remove feelings of annoyance or spite and instead inspire and motivate users.

A simple "I was paying attention" button, alongside the nudge dismissal button, could be used to feed a Machine Learning model labelled data, that would later become more refined and thus more personalized. This would also help calibrate attention thresholds and thus minimize user frustration.

10.2 Advanced Eye-Tracking

Also based on user feedback, advancements in eye-tracking algorithms could further increase the efficacy of pin-pointing the potential sources of attentional lapses by examining the region on the screen where the attention might have gotten stuck, or paragraphs revisited one too many times by the user. Such a feature could enable content-aware nudges or (probably more preferably) content-aware explanations, recontextualizations, or even a voice chat assistant with the initial prompt of "I don't understand this part: ...". Given the current state of webcam resolutions and computer vision techniques, such real-time implementation of said features is increasingly feasible.

Appendix: Barplots

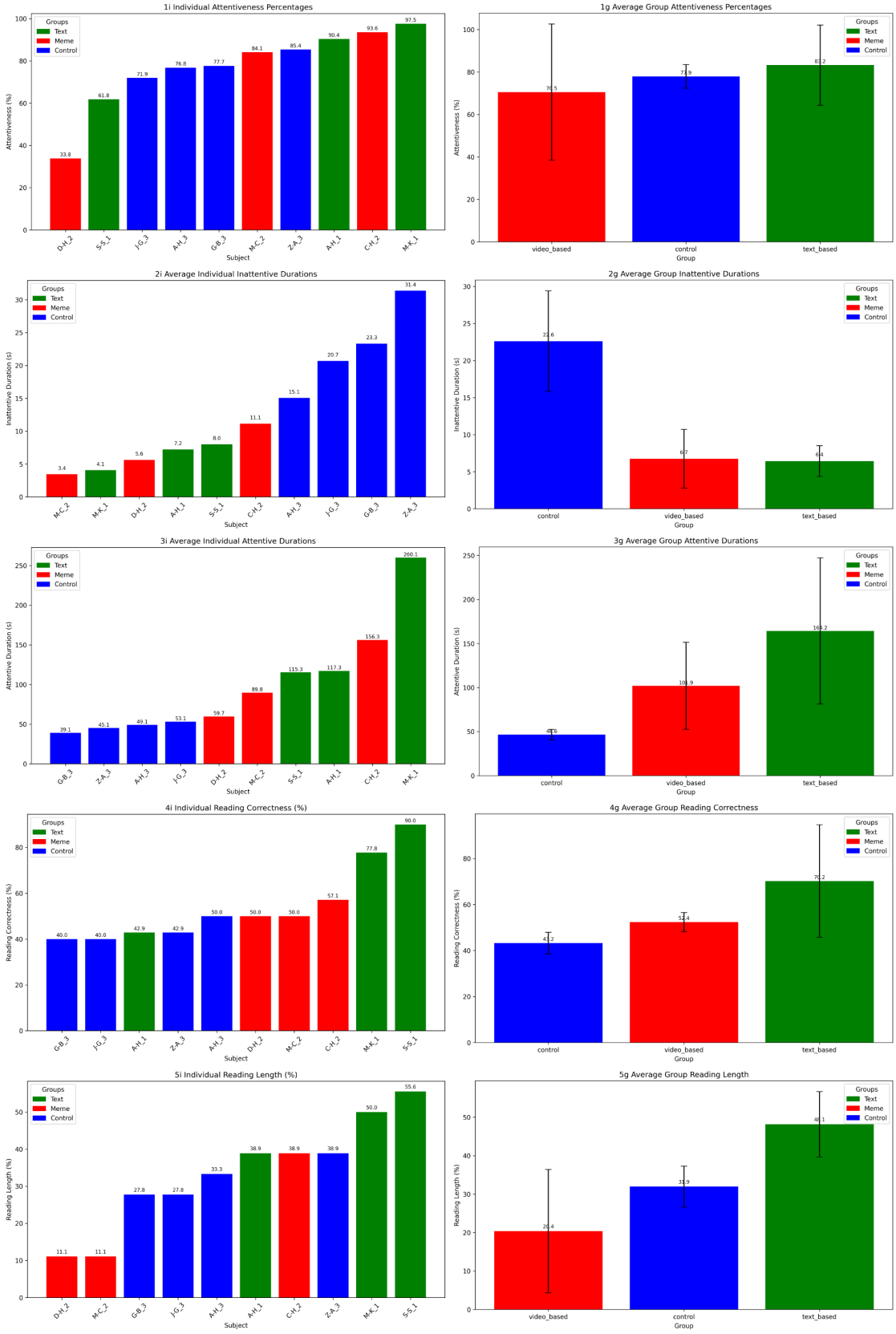

Fig. 3: Overview of attentiveness and reading metrics across individual and group conditions.

Appendix: Attention Score Lineplots

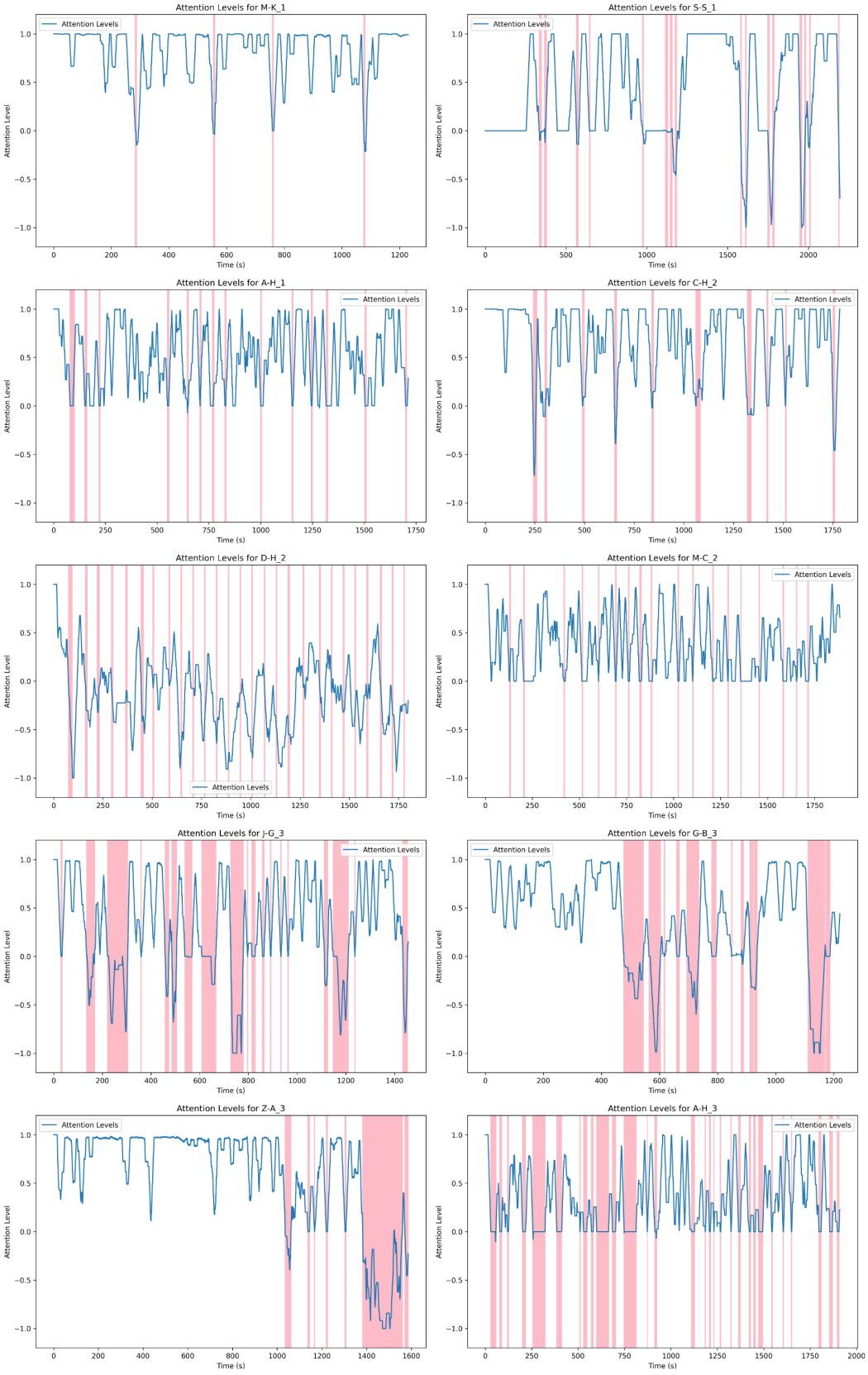

Fig. 4: All collected session data with users. Group are represented with a numeric suffix, 1=Text, 2=Meme, 3=Control

Appendix: Pre- and Post-Experiment Questionnaire

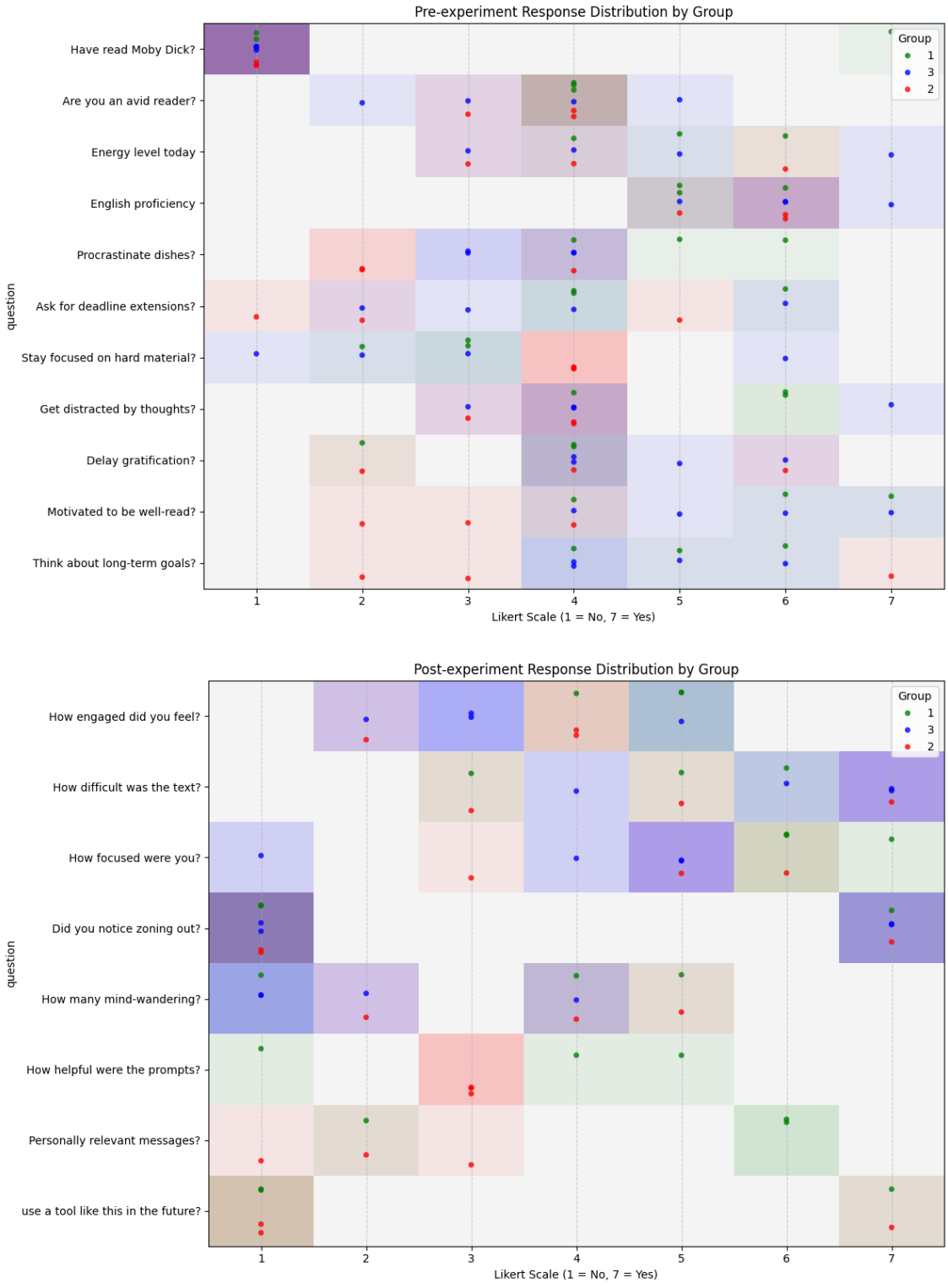

Fig. 5: All qualitative questions and answers of participants. Groups are represented with a numeric suffix: 1=Text, 2=Meme, 3=Control.

Appendix: Participant Responses

User Responses by Group

Text Nudging Group (1)

A H System was too sensitive, it bothered me. I was wearing glasses, that could be the issue. Some parameters need be changed; text was hard to interpret, I had to re-read lines. It didn't feel personal enough these attention prompts, the system would need to get to know me better over a longer period of time. I liked the questionnaire a lot after the reading! Better than the system itself.

M K I feel like this program works great. I did pay attention, but for instance my roommate came in during the reading and I didn't look for a few minutes and it immediately told me to pay attention more. I think this is especially great if it comes to exam preparations where every second counts. Props for the dev team!

S S I had a sense of urgency, almost out of spite, to really pay attention after the nudge prompts. The program used nudges a bit too often, even while I was reading. It would be great if it had a function where, if I didn't understand something, it would automatically show me a helpful bubble explaining the part that I didn't get.

Meme Nudging Group (2)

C H I didn't like it because the videos were not funny, and if I read I want to read and don't want to see videos. The reading was very difficult.

D H It was annoying that, although I did pay attention, the program nudged me to continue. It disturbed my attention, because the videos made me laugh and actually took me longer to get back to the material.

M C In a third of the cases it was accurate; other times it was quite bothering. I became resentful and spiteful and I paid extra attention so that the system would not bother me as much. Sometimes, even though I moved my eyes and "read", I didn't comprehend. So it is complicated. It is not just about eye movements.

Appendix: Stimulus Construction

Example Goal-Setting Response: The following is an example of user goals. These were processed with the same LLM model to remove phrasing characteristics and other PID's. *List 4–5 medium- or long-term goals you have in life:*

> Personally, I seek meaningful relationships, marriage, and fatherhood, embracing the role of *pater familias*. I also have a deep desire to travel, reconnect with my childhood experiences, and immerse myself in diverse cultures—particularly in places like Australia and remote islands—to witness the world in its raw, natural beauty.

How would becoming more literate or well-read help you achieve these goals?

> I want to live up to the literary example set by my parents and become as well-read as my brother. I recognize that even books I disagree with could broaden my worldview in the long run.

Nudging Message Generation A local LLaMA 3.1 8B model was used to generate 10–15 personalized nudging messages per participant, based on the life goals and contextual framing they provided during the pre-experiment phase.

Standardized Prompt:

> Generate 20 short nudging messages based on this information... Don't be too motivating... focus on goal–content connection.

The model's output was manually reviewed and slightly curated by Researcher as he saw fit to ensure consistent formatting and content clarity or to remove nonsensical responses. Despite local fine-tuning, occasional deviations in output structure made human oversight quite necessary.

Appendix: Reflection

This project was among the most difficult projects of my university years. Despite a strong background in programming – including extensive tutoring experience – there were real technical problems I encountered while writing the model and the experiment programs.
One such problem that I had to circumvent is the abysmal state of video (and general media) playback in a Python environment. Many libraries either lacked essential features or were not thread-safe, meaning that one process cannot start another to run alongside with it, as in the case of these multimedia libraries the closing of the child process (the playback) would unintentionally close the parent as well, thus shutting the entire experiment down. It would be an understatement if I said I spent at least a week reverse engineering the libraries, trying to make them thread-safe, but in the end I used a VLC module instead (that was also unbearably buggy but at least safe to use in a delicate setting).
Another problem was the search for participants. Asking volunteers to engage in a 30-minute reading task – with minimal incentives: a croissant and a drink – truly limited the number of participants. Though this trade-off arguably contributed to higher data quality. Larger incentives might have introduced motivational bias. In this case, fewer but more attentive participants likely improved internal validity.

References

1. Anh, N.T.L., Bach, N.G., Tu, N.T.T., Kamioka, E., Tan, P.X.: Svd-based mind-wandering prediction from facial videos in online learning. Journal of Imaging **10**(5), 97 (2024). https://doi.org/10.3390/jimaging10050097
2. Baldassarri, S., Hupont, I., Abadía, D., Cerezo, E.: Affective-aware tutoring platform for interactive digital television. Multimedia Tools and Applications **74**(9), 3183–3206 (2015)
3. Bradbury, N.A.: Attention span during lectures: 8 seconds, 10 minutes, or more? Advances in Physiology Education **40**(4), 509–513 (2016). https://doi.org/10.1152/advan.00109.2016
4. Broadbent, D.E.: Perception and communication. Pergamon Press (1958). https://doi.org/10.1037/10037-000, p. 42
5. Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O.: Dopamine in motivational control: rewarding, aversive, and alerting. Neuron **68**(5), 815–834 (2010). https://doi.org/10.1016/j.neuron.2010.11.022
6. Buono, P., De Carolis, B., D'Errico, F., Macchiarulo, N., Palestra, G.: Assessing student engagement from facial behavior in on-line learning. Multimedia Tools and Applications **82**(9), 12859–12877 (2023)
7. Darnell, D.K., Krieg, P.A.: Student engagement, assessed using heart rate, shows no reset following active learning sessions in lectures. PLoS ONE **14**(12), e0225709 (2019). https://doi.org/10.1371/journal.pone.0225709, open access under Creative Commons Attribution License
8. Digital Learning Institute: Unveiling the power of cognitive science in digital learning. https://www.digitallearninginstitute.com/blog/unveiling-the-power-of-cognitive-science-in-digital-learning (2025), accessed: 2025-05-29
9. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learning and Instruction **29**, 153–170 (2014)
10. Engle, R.W., Kane, M.J.: Executive attention, working memory capacity, and a two-factor theory of cognitive control. In: Ross, B.H. (ed.) The Psychology of Learning and Motivation, vol. 44, pp. 145–199. Academic Press (2003). https://doi.org/10.1016/S0079-7421(03)44005-X

11. García-Méndez, S., de Arriba-Pérez, F., Somoza-López, M.d.C.: A review on the use of large language models as virtual tutors. Science & Education **34**, 877–892 (2025). https://doi.org/10.1007/s11191-024-00530-2, published 18 May 2024, Accepted 26 April 2024
12. Gerappa, U.: attention. https://github.com/gerappa01/attention (2025), accessed: 2025-07-17
13. Gupta, S., Kumar, P., Tekchandani, R.K.: Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models. Multimedia Tools and Applications **82**(8), 11365–11394 (2023)
14. Handayani, D., Yaacob, H., Rahman, A.W.A., Sediono, W., Shah, A.: Systematic review of computational modeling of mood and emotion. In: Proceedings of the 5th International Conference on Information and Communication Technology for the Muslim World (ICT4M). pp. 1–5. IEEE, Kuching, Malaysia (2014). https://doi.org/10.1109/ICT4M.2014.7020587
15. Kane, M.J., Brown, L.H., McVay, R.C., Silvia, P.J., Myin-Germeys, I., Kwapil, T.R.: For whom the mind wanders, and when: an experience-sampling study of working memory and executive control in daily life. Psychological Science **18**(7), 614–621 (2007). https://doi.org/10.1111/j.1467-9280.2007.01948.x
16. Van der Kleij, F.M., Feskens, R.C.W., Eggen, T.J.H.M.: Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. Review of Educational Research **85**(4), 475–511 (2015)
17. Lachter, J., Forster, K.I., Ruthruff, E.: Forty-five years after broadbent (1958): Still no identification without attention. Psychological Review **111**(4), 880–913 (2004). https://doi.org/10.1037/0033-295X.111.4.880
18. Li, Q., Fu, L., Zhang, W., Chen, X., Yu, J., Xia, W., Zhang, W., Tang, R., Yu, Y.: Adapting large language models for education: Foundational capabilities, potentials, and challenges. arXiv preprint arXiv:2401.08664 **9** (2023)
19. Linnenbrink-Garcia, L., Patall, E.A., Pekrun, R.: Adaptive motivation and emotion in education: Research and principles for instructional design. Policy Insights from the Behavioral and Brain Sciences **3**(2), 228–236 (2016)
20. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for perceiving and processing reality. In: Proceedings of the Third Workshop on Computer Vision for AR/VR at IEEE CVPR (2019), https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf
21. Onah, D.F.O., Sinclair, J., Boyatt, R.: Dropout rates of massive open online courses: Behavioural patterns. In: EDULEARN14 Proceedings. pp. 5825–5834. IATED, Barcelona, Spain (2014), 6th International Conference on Education and New Learning Technologies
22. Paprocki, R., Lenskiy, A.: What does eye-blink rate variability dynamics tell us about cognitive performance? Frontiers in human neuroscience **11**, 620 (2017)
23. Posner, M.I., Petersen, S.E.: The attention system of the human brain. Annual Review of Neuroscience **13**, 25–42 (1990). https://doi.org/10.1146/annurev.ne.13.030190.000325
24. Riby, L.M., Marr, L., Barron-Millar, L., Greer, J., Hamilton, C.J., McGann, D., Smallwood, J.: Elevated blink rates predict mind wandering: Dopaminergic insights into attention and task focus. Journal of Integrative Neuroscience **24**(3), 26508 (2025)
25. Risko, E.F., Anderson, N., Sarwal, A., Engelhardt, M., Kingstone, A.: Everyday attention: variation in mind wandering and memory in a lecture. Applied Cognitive Psychology **26**(2), 234–242 (2012). https://doi.org/10.1002/acp.1814
26. Robal, T., Zhao, Y., Lofi, C., Hauff, C.: Webcam-based attention tracking in online learning: A feasibility study. In: Proceedings. pp. 189–197 (2018). https://doi.org/10.1145/3172944.3172987
27. Ross, M., Graves, C., Campbell, J., Kim, J.: Using support vector machines to classify student attentiveness for the development of personalized learning systems. In: 12th International Conference on Machine Learning and Applications (ICMLA). vol. 1, pp. 325–328 (2013). https://doi.org/10.1109/ICMLA.2013.66

28. Saxena, S., Fink, L.K., Lange, E.B.: Deep learning models for webcam eye tracking in online experiments. Behavior Research Methods **56**, 3487–3503 (2024). https://doi.org/10.3758/s13428-023-02190-6, published August 22, 2023; Issue Date: June 2024
29. Schultz, W.: Phasic dopamine responses: a neural substrate of reward prediction error. Nature Reviews Neuroscience **12**(11), 755–768 (2011). https://doi.org/10.1038/nrn3109, available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3032992/
30. Szpunar, K.K., Moulton, S.T., Schacter, D.L.: Mind wandering and education: from the classroom to online learning. Frontiers in Psychology **4**, 495 (2013). https://doi.org/10.3389/fpsyg.2013.00495, section: Perception Science, Research Topic: Towards a psychological and neuroscientific account of the wandering mind
31. Treisman, A.M.: Selective attention in man. British Medical Bulletin **20**, 12–16 (1964). https://doi.org/10.1093/oxfordjournals.bmb.a070274
32. Waldorf Today: In china, classroom cameras scan student faces for emotion—stoking fears of new form of state monitoring (Apr 2019), https://www.waldorftoday.com/2019/04/in-china-classroom-cameras-scan-student-faces-for-emotion-stoking-fears-of-new-form-of-state-monitoring/
33. WebGazer Team: Webgazer.js: Eye tracking with javascript in the browser. https://webgazer.cs.brown.edu/ (2025), accessed: 2025-07-13
34. Yuvaraj, R., Mittal, R., Prince, A.A., Huang, J.S.: Affective computing for learning in education: A systematic review and bibliometric analysis. Education Sciences **15**(1) (2025). https://doi.org/10.3390/educsci15010065