This is the author's version of the article. To appear in an IEEE ISMAR conference.

PersoNo: Personalised Notification Urgency Classifier in Mixed Reality

Jingyao Zheng*
The Hong Kong Polytechnic University
Haodi Weng†
The Hong Kong Polytechnic University
Xian Wang‡
The Hong Kong Polytechnic University
Chengbin Cui§
The Hong Kong Polytechnic University
Sven Mayer¶
TU Dortmund University
Chi-lok Tai‖
The Hong Kong Polytechnic University
Lik-Hang Lee**
The Hong Kong Polytechnic University

Abstract

Mixed Reality (MR) is increasingly integrated into daily life, providing enhanced capabilities across various domains. However, users face growing notification streams that disrupt their immersive experience. We present *PersoNo*, a personalised notification urgency classifier for MR that intelligently classifies notifications based on individual user preferences. Through a user study (N=18), we created the first MR notification dataset containing both self-labelled and interaction-based data across activities with varying cognitive demands. Our thematic analysis revealed that, unlike in mobiles, the activity context is equally important as the content and the sender in determining notification urgency in MR. Leveraging these insights, we developed *PersoNo* using large language models that analyse users' replying behaviour patterns. Our multi-agent approach achieved 81.5% accuracy and significantly reduced false negative rates (0.381) compared to baseline models. *PersoNo* has the potential not only to reduce unnecessary interruptions but also to offer users understanding and control of the system, adhering to Human-Centered Artificial Intelligence design principles.

Index Terms: Mixed Reality, Notification Classifier, Human Centered Artificial Intelligence.

1 Introduction

Mixed Reality (MR) environments are increasingly integrated into daily life, blending digital information with physical surroundings. In this paper, we treated MR as synonymous with Augmented Reality: virtual objects integrated into the real world [57]. It enhances human capabilities across manufacturing [14] and education [29]. However, to avoid losing touch with reality, users face growing notification streams in MR, which present unique challenges as they distract users from their immersive experience with no task-related information. Virtual Reality (VR) studies have discussed similar concerns of breaking the immersive experience and emphasised the importance of notifications [13, 18, 52]. Yet isolating users from notifications induces anxiety and disconnection [42]. This contradiction underscores the need for intelligent MR notification classifiers to filter notifications and cause less disruption appropriately.

Human-computer interaction (HCI) researchers have studied how to balance user attention and interruptions in mobile settings for a long time. Prior work has revealed that the content of a message is a primary factor determining its urgency and a user's likelihood of responding [25, 33]. In addition, users' recent interactions with senders allow them to make accurate notification speculations [6], indicating the sender is pivotal in determining users' receptivity to the notifications. Prior approaches [33, 35] have leveraged such insights to develop intelligent context-aware smartphone notification systems. However, mobile users can easily distance themselves from interruptions by putting their phones away. In contrast, to truly escape notifications in MR, users must completely remove their head-mounted devices (HMDs), which not only breaks the immersive experience but also eliminates access to all the augmented capabilities that make MR valuable for their work.

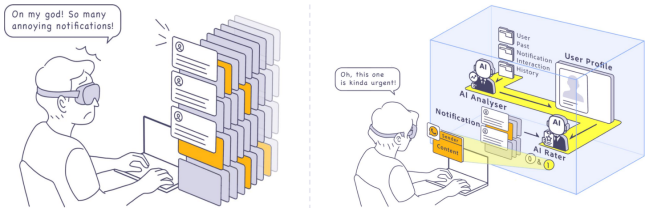

Figure 1: **Left**: Users are overwhelmed by MR notifications during work, with only a few urgent notifications. **Right**: With **PersoNo**, only urgent notifications are pushed to the user. The box illustrates the system workflow—the AI analyser processes the user's past notification interaction history to generate user profiles, which AI raters then use to evaluate upcoming notifications.

Unlike mobile notification researchers, MR/VR researchers primarily focused on notification design and multimodal approaches [13, 31, 44, 51]. Limited research exists on intelligent notification management [7, 28], yet we argue it is crucial for alleviating user distraction. To date, no established dataset or framework captures how users handle incoming notifications in MR, leaving developers without datasets for developing MR notification classifiers. This knowledge gap, coupled with the practical importance of minimising user distraction in MR, motivates our research.

In this work, we aim to manage notifications intelligently to reduce user distraction while maintaining important notification updates. We address two research questions that drive our study and facilitate the development of the Personalised Notification Urgency Classifier in MR (PersoNo): **RQ1: How do users behave and respond to notifications in MR?** We seek to understand the human side of the problem: when an MR user receives a notification, what factors influence whether they attend to it or ignore it? Would the variables be the same as those in mobile notification interaction? These responses inform the key variables to be considered in developing MR notification classifiers, providing critical insights into contextual and user-specific factors. Regarding the second research question, we draw upon previous notification classifier research [7, 11, 25], which demonstrated that personalised models trained exclusively on individual user data outperform gen-

*e-mail: jingyao.zheng@connect.polyu.hk
†e-mail: wengvictor5@gmail.com
‡e-mail: xiann.wang@connect.polyu.hk
§e-mail: chengbin.cui@connect.polyu.hk
¶e-mail: info@sven-mayer.com
‖e-mail: andy.tai@cpce-polyu.edu.hk
**Corresponding Author. e-mail: lik-hang.lee@polyu.edu.hk.

eral models trained on aggregated multi-user datasets. Based on these findings, we formulate **RQ2: How can we automatically classify the urgency of MR notifications in a personalised manner?** which encompasses three main dimensions for developing a classifier: Data, Context, and Algorithm.

To address these questions, we conducted a detailed study and developed a solution with three key contributions, corresponding to three PersoNo essential elements (Data, Context and Algorithm): **(1)** We created a **new MR notification dataset** through a user study (N=18) where participants wore MR headsets while experiencing everyday tasks and received messages. We collected objective and subjective data through self-labelling (users' rating notification urgency) and by tracking actual response behaviours. With this first-of-its-kind dataset ($N = 18 \times 198$), we demonstrated that self-labelling offers a convenient alternative to activity-based data collection for future PersoNo deployment, yielding comparable classifier performance. **(2)** We analysed **users' replying behaviour patterns regarding MR notifications**. Our findings revealed certain patterns consistent with mobile research. For example, message content emerged as a critical factor when determining whether to attend to notifications [25]. However, we also discovered MR-specific insights. Notably, activities were reported with similar frequency as content when users described their behavioural patterns. **(3)** Our proposed **PersoNo** algorithm leverages Large Language Models (LLMs) and its classifier could accurately predict notification urgency by analysing users' replying behaviour patterns in small notification datasets.

2 RELATED WORK

Digital Notifications Researchers have examined notifications on smartphones and other personal devices. Mobile users receive dozens of notifications per day (around 63.5 on average), primarily from messaging and email, and typically attend to them within minutes due to social pressures [41]. While frequent alerts can induce stress or a sense of interruption, users also report feeling more connected when messaging notifications keep them aware of social updates [5]. Notably, complete avoidance of notifications is not a viable solution; experiments disabling push alerts found that users experienced anxiety and isolation without these ambient cues [42]. This underscores the need to manage rather than eliminate digital interruptions. Prior work identified key factors that determine which notifications users deem urgent or worthy of immediate response. The content of the message is consistently found to be a primary influence on perceived urgency and responsiveness [25, 33]. For example, critical or work-related content demands quicker attention than trivial updates. The sender is another pivotal factor: Chang et al. [6] observed that users often speculate about who a notification is from, and recent interactions with a sender strongly influence whether they will check the alert immediately. Other contexts also influence users' receptivity. These contexts include location [33, 39], time of day [46, 54] and activity context [2, 34]. These insights informed the design of intelligent notification management systems that attempt to filter or rank alerts by importance.

While existing notification research provides valuable insights for users' receptivity to mobile notifications, these findings may not translate to MR directly. Unlike mobile notifications that users can physically distance themselves from, MR notifications are inherently more invasive due to their immersive presentation within the user's field of view. Users must either endure disruptions or remove headsets entirely, sacrificing all augmented capabilities. It might potentially increase users' notification fatigue. This fundamental difference necessitates specialised approaches for MR notification management. Our work investigates how established factors influencing notification receptivity manifest differently in MR, and develops personalised intelligent systems based on the most significant contexts.

Personalised Notification Classifier Research on intelligent notification systems has explored multiple strategies: opportune time predicting [7, 36, 47, 55] and notification management [25, 32]. Building upon this foundational research, subsequent studies [7, 11, 25] have compared models trained on personal and generic data, revealing a consistent pattern: personal data enables higher accuracy in notification classification. This raises a critical research challenge: how to construct an intelligent management system with limited training data.

Prior works developed intelligent notification management systems from both subjective [7, 34] and objective experience [32, 40]. Inspired by this, we compared two data collection methods: self-labelling and interaction, both previously used in message classification research [11, 7]. They were included in our study, as each offers distinct advantages. Previous interaction-based data collection typically required several weeks to gather sufficient classifier training data. For example, Mehrotra et al. [32] needed a 15-day experiment to collect mobile notifications, while Pielot et al. [40] spent an average of four weeks gathering data. In contrast, self-labelling allows notifications to be categorised within a much shorter timeframe. This efficiency could significantly enhance user acceptance of classifier applications, as users typically prefer applications that are ready for use shortly after deployment [30]. However, prior work [10] indicates that self-reporting and actual behaviours are only weakly correlated. It suggests that self-labelled data might not be reliable for training personalised classifiers. Our study, therefore, aims to compare the accuracy of classifiers trained on both types of data to determine whether self-labelled data can effectively substitute for interaction-based data in this context.

Notifications in Mixed Reality As computing extends into immersive environments like VR, notification management faces new challenges. In VR, users can become so engrossed that they miss critical external messages, leading to frustration when important information is delayed [18]. To address this challenge, numerous VR researchers have investigated optimal placement strategies to ensure user visibility and attention [18, 52, 19]. Besides the visual cues, Ghosh et al. [13] explored multiple modalities for VR notifications. Among visual, aural, and haptic notifications, haptic ones were the least effective, a finding that aligns with subsequent research by George et al. [12].

While VR notification research has made considerable progress, MR introduces further complexity as digital information overlays the real world rather than replacing it. Prior works have explored the effectiveness of notifications across different multimodalities [31, 9]. From a visual perspective, Rzayev et al. [51] demonstrated that notification positioning significantly impacts user perception, with proper alignment crucial for minimising distraction while maintaining awareness. Notably, Li et al.[26] developed a computational framework that predicts virtual element noticeability by analysing visual saliency patterns to anticipate when users detect element changes.

While existing research in MR and VR has predominantly focused on notification design elements like placement, modality, and visual appearance, there remains a significant gap in an intelligent MR notification management system. In this context, early work by Orlosky et al. [37] showed that using see-through HMDs to relay mobile notifications can increase message awareness with minimal performance impact compared to checking handheld phones. However, this advantage diminishes in high cognitive load situations where users exhibit varying receptivity to interruptions; only HMDs equipped with intelligent notification systems that adapt to users' cognitive states and contextual preferences can truly deliver benefits without compromising task performance, such as the adaptive MR user interfaces based on users' cognitive load [28]. Our research first investigates the most significant contexts affecting MR users' receptivity to notifications, leveraging participant-reported

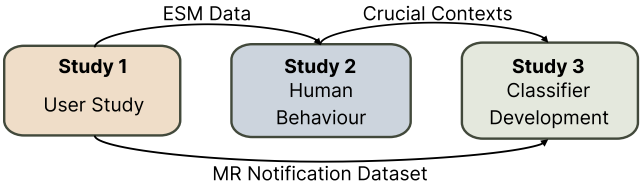

Figure 2: The research approach we took for our investigation.

contexts to develop an MR notification management system that enhances user experience. More broadly, our work focuses on mitigating distraction in MR. While previous research has explored distraction reduction in MR/VR generally [48, 63, 49], our work makes a distinct contribution by specifically focusing on distraction mitigation through intelligent notification management.

3 Research Approach

Our research addresses interaction challenges in MR notification systems through a three-stage approach depicted in Figure 2. Each study addressed one core PersoNo classifier component: Study 1 focused on Data, Study 2 on Context, and Study 3 on Algorithm. Study 1 collected an MR notification dataset through a user study, complemented by the Experience Sampling Method (ESM) [17] to capture participants' behavioural patterns through immediate self-reports. This facilitates subsequent analysis of optimal data collection methods (self-labelled or interaction-based) for future PersoNo deployment. The ESM data directly informed Study 2, where we analysed human behaviour patterns regarding notification interaction in various contexts. This identifies the key variables PersoNo should consider when classifying notification urgency levels. Building on these insights, Study 3 leveraged the MR notification datasets and crucial contexts to develop an intelligent MR notification management system, PersoNo.

Our work identifies crucial contextual variables that influence notification receptivity in MR. It provides both empirical insights into user behaviour and practical solutions for reducing notification distraction while maintaining awareness of important information. This integrated approach bridges human-centred research with advanced Artificial Intelligence (AI) techniques to address a significant usability challenge in emerging MR interfaces.

4 Study 1: MR Notification Data Collection

To the best of our knowledge, the field lacks a comprehensive MR notification interaction dataset. To build an MR notification classifier, our research requires a user study to collect MR notification data. This collection serves two key purposes: analysing how participants respond to notifications in MR and developing an effective notification classification system.

To collect both the subjective and objective notification dataset, two phases (self-label phase and interaction phase) were conducted in a counter-balanced order. In the self-label phase, participants assessed the urgency levels of 90 randomly selected notifications from our dataset (details in Section 4.1.1), given the message content and the senders. For the interaction-based data collection phase, participants engaged in three MR activities, each consisting of two ten-minute sessions. Between sessions, there was at least a one-minute break. Our system recorded participants' behavioural responses and classified notifications as 'non-urgent' when participants either ignored or dismissed them, and as 'urgent' when participants actively chose to respond within 30 seconds.

Our approach classified notification urgency into two categories: *urgent* and *non-urgent*. This binary classification builds upon the work of Weber et al. [61], who initially identified four notification clusters (*C1*, *C2*, *C3*, and *C4*) in daily interactions. Their research revealed that only *C1* notifications demanded immediate user attention, while *C2*, *C3*, and *C4* could be addressed at the user's convenience. Thus, we used a binary classification in our study based on whether immediate attention is required, which also aligns with the previous notification research design [33]. We define the *urgent* notifications as those that require replies within 30 seconds, and *non-urgent* notifications which did not have this time constraint.

We initially hypothesised that considering only two key variables (content and sender) could achieve high prediction accuracy, as previous research [6, 25, 34] suggested that these variables alone could yield reasonable results. Additional factors incorporated into the further analysis include the activities in which users were involved and their established messaging reply habits, as determined through the following thematic analysis. Our contribution in this section lies in the construction of an MR notification dataset based on users' interaction behaviour during the MR activities ($N = 18 \times 108$, comprising 18 participants with 108 notification data points per participant) and a self-labelled dataset ($N = 18 \times 90$).

4.1 Mixed Reality Notification

4.1.1 Notification Dataset

Our study focused exclusively on instant messaging (IM) notifications, as mobile IM messages are anticipated to become a fundamental MR component [24]. We selected WhatsApp as the application source due to its widespread use in our region. For the MR activity data, we carefully balanced the quantity of data with notification frequency. While aiming to maximise data collection, we avoided pushing notifications too frequently to prevent user annoyance, establishing a reasonable notification interval (See Section 4.4). In total, we collected 108 notification instances during the MR activities and separated them into training and testing datasets (90 and 18, respectively; more details in Subection 6.1). Our approach follows established methodological practices in notification research [7, 25] that separate activity-generated data into distinct training and testing datasets. To ensure equivalent training sets across both collection methods, we also gathered 90 self-labelled data, resulting in 198 notification data.

To protect participant privacy, we used Python scripts to randomly extract 198 data from the online Mobile Text Dataset (*mobile_train.txt*) [59]. This dataset is grounded in real-world mobile user behaviour. Originally, the dataset only contained the message content. To emulate the real-world experience, we assigned sender placeholders for each notification, such as *friend 1* and *friend 2*. Similar to the previous work [52], we collected the names of participants' friends and supervisors to replace the sender placeholders before the study and used these names as message senders during the experiment. To further enhance the realistic experience, we also incorporated group messages at proportions similar to those reported in Pielot et al.'s work [43] (40 group messages and 158 messages).

4.1.2 Notification Interaction

We designed notification interaction to mirror actual behaviours: users can ignore, actively dismiss, or respond to notifications (see

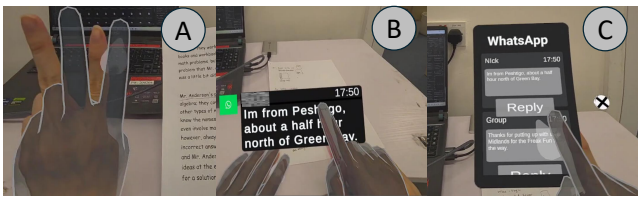

Figure 3: Examples of Notification Interaction. (A) demonstrates the gesture for dismissing notifications in MR. (B) illustrates the tapping interaction to access the notification panel displayed in (C), which features the "Reply" button that initiates the response workflow.

Figure 3). During the activities, participants may ignore or dismiss *non-urgent* notifications. If they did not actively dismiss or respond within 20 seconds, notifications were automatically dismissed and stored in the notification panel (see Figure 3C), aligning with previous research [52]. Additionally, participants had the option to manually dismiss notifications using a specific gesture (see Figure 3A). A response was required for notifications deemed urgent by the participants. Unlike the prior studies [7, 52] that only allowed quick responses through simple controller presses, our study required participants to take additional steps to reply due to the common practice where users typically respond to notifications manually rather than using the smart reply [22].

Participants needed to open the app by either tapping the notification or gesturing to respond to notifications. We omitted the message-typing step and utilised the "Reply" button click to simulate the reply process (see Figure 3C) to streamline the process and avoiding fatigue and dizziness caused by longer study duration.

4.1.3 Notification Display

We adopted a notification user interface design (see Figure 4D-F) similar to previous work [7, 52], which displays the sender, an image of the application source, and the content. All notifications were placed within the participants' field of view and designed to be easily noticeable. Specifically, notifications were placed in the bottom centre of the user's field of view, as prior work [53] showed this position improves comprehension and reduces distraction while sitting and Plabst et al. [45] found subtitles provide higher comprehension and noticeability than heads-up displays. We positioned notifications 0.25 meters from users, closer than the Quest 3's focal distance of over 1 meter, to accommodate table-based MR tasks. Greater distances risked users reaching through the table when tapping, potentially causing injury. Overall, notifications were placed 0.25 meters away and angled 25° below the user's line of sight.

4.2 Procedure

The entire study lasted approximately two hours. Upon arrival, participants were welcomed and provided with the information sheet detailing the study's purpose. Then, they signed a consent form and completed a demographic questionnaire. As mentioned earlier, before conducting the formal study, we asked their friends and bosses/supervisors for a few names. This information was filled in our notification dataset to replace the placeholders, such as *Friend 1* and *Supervisor*. Then, we utilised Python scripts to randomly separate the data into two groups: self-labelled notifications (90) and MR activity notifications (108). Following this, we counterbalanced the order of the two data collection methods. During the self-label phase, participants were asked to carefully read the notification content and the sender, and then rate the urgency levels of each notification according to their preferences and daily habits. The self-labelling session was conducted on a laptop. Regarding the MR activity part, we conducted an introductory session using slides to outline basic interactions with notifications and the primary tasks for each activity. Additionally, we developed an introductory VR scene that allowed participants to familiarise themselves with the notification interactions in a controlled environment. Once they were confident in the interaction, the MR activities were conducted in a Latin Square order to alleviate the carryover effects.

4.3 Participants

Our university's ethics board approved the study. We compensated participants through course credit or payment at the local minimum wage. We recruited 18 participants (5 males and 13 females) from our universities, aged from 18 to 27 years (M = 22.61, SD = 2.63). All participants had either normal or corrected-to-normal vision and were able to view notification details clearly. To measure participants' familiarity with MR, we used a 5-point Likert scale, where 1 indicated very inexperienced ("I have only used MR once or twice before, if at all") and 5 indicated very experienced ("I use MR several times a month"). Results showed that participants were generally unfamiliar with MR (M = 2.5, SD = 1.12).

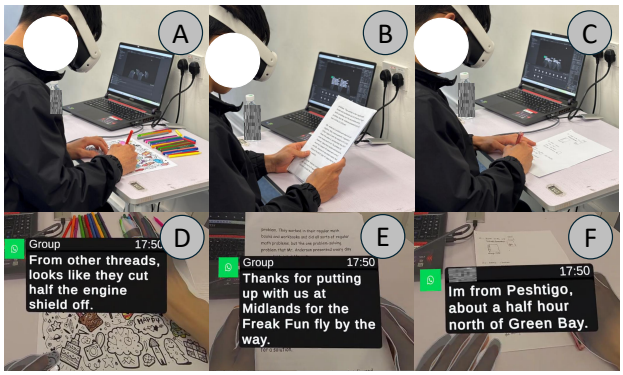

Figure 4: Examples of Mixed Reality Activities. (A) shows the *Doodling* Activity. (B) illustrates the *Reading and Comprehension* Activity. (C) presents the *Brainstorm* Activity. (D), and (E), (F) show the examples of notifications received during the activity.

We also collected data regarding participants' mobile IM notification preferences. Participants were asked to select contexts and the most important one affecting their receptivity to mobile IM notifications, including Notification Content, Sender, Application Source, Cognitive Load, Time, Location, Mood, Activity, and others [5, 34]. 14 participants identified Notification Content as a significant context, while 7 selected Sender and 9 chose Activity. Regarding the most influential factor, Notification Content emerged as the most influential factor, reported by 10 out of 18 participants, followed by Sender (N=3) and Application Source (N=3). Individual contextual factors such as Mood (N=1) and Activity (N=1) showed notable variability among participants. These findings align with previous research indicating that most participants consider notification content the most influential contextual factor [25].

4.4 Experiment Design

Mixed Reality Activities Users engaged in three MR activities: *Doodling*, *Brainstorming*, and *Reading and Comprehension* (see Figure 4 (A-C)). These activities require varying levels of cognitive load, ranging from low to high [28]. It created a realistic testing environment where users encountered notifications across different mental states and activity types, similar to daily life experiences. During the *Doodling* activity, participants were provided with five different plain graffiti drawings and coloured pens. They were free to choose one drawing and doodle without specific requirements. For the *Brainstorming* activity, participants focused on designing future notifications in MR. Following Rietzschel et al.'s guideline [50], participants wrote their ideas on A4-sized sheets of paper and were encouraged to think creatively without concerns about feasibility. For example, they were prompted to consider multi-modal notifications beyond visual elements, including haptic feedback and taste. The *Reading and Comprehension* materials and questions were sourced from easyCMB [1], a resource widely used in previous research [8, 21]. Participants were instructed to read as quickly as possible while ensuring accuracy in their answers.

Data Collection The primary objective of our experiment was to collect notification data categorised by urgency levels. The study consisted of two parts: a self-label session and an interaction session. In the self-label session, participants labelled the urgency levels of each notification (90 notifications) in a CSV file. In the interaction session, participants engaged in three MR activities, with

each activity divided into two 10-minute sessions. For each activity session, participants performed a primary task (the MR activity) while simultaneously handling a secondary task: reading notifications carefully and deciding whether to reply to messages. During each 10-minute activity session, 18 notifications were sent to participants at random intervals ranging from 20 to 32 seconds (cf. [52]). Our script automatically classified notifications based on response time: if a notification received a reply within 30 seconds, it was labelled as *urgent* (1); otherwise, it was labelled as *non-urgent* (0). Our system yielded two distinct datasets for analysis in Section 6. Dataset 1 comprises information about the Sender, Content, and Urgency level, while Dataset 2 expands upon this by including Sender, Content, Urgency level, and Activity context. To further protect participant privacy, we parsed their friend and supervisor names to generic placeholders (e.g., *Friend 1* and *Supervisor*) throughout the dataset. To gain deeper insights into participants' decision-making processes, we applied the Experience Sampling Method (ESM) [17]. During the user study, at least one researcher periodically asked them to explain their notification response patterns. For example, a common prompt was: "Could you please explain why you replied to your friend's last message?". These qualitative insights were summarised and documented. It was incorporated into our analysis (see section 5).

Additionally, we measured participants' cognitive load during the study. While Lindlbauer et al. [28] employed the three MR activities as requiring different cognitive loads in application scenarios, they did not empirically demonstrate these differences. To address this gap, we administered the NASA-TLX questionnaire [15], which provided a standardised assessment of participants' perceived cognitive workload across the different activities.

Apparatus We utilised a Windows 11 laptop with an NVIDIA GeForce RTX 4080, an Intel Core i9-13980HX processor, and 32GB RAM, which connected to the Meta Quest 3 headset. This configuration supported the pass-through functionality required for our MR environment. The software application was developed and implemented using Unity version 2022.3.22f1.

4.5 Results

The NASA-TLX analysis revealed significant differences in cognitive workload across the three conditions (Doodling, Brainstorming, and Reading). Shapiro-Wilk tests confirmed normal distribution only for Performance across activities (Doodling: $p = .283$, Brainstorm: $p = .316$, Reading: $p = .243$) with no significant differences found via one-way repeated measures ANOVA ($p = .190$). Friedman tests for non-normally distributed dimensions revealed significant differences in Mental Demand ($\chi^2(2) = 13.82, p = .001$), Temporal Demand ($\chi^2(2) = 9.97, p = .007$), Effort ($\chi^2(2) = 8.27, p = .016$), and Frustration ($\chi^2(2) = 10.68, p = .005$). Post-hoc analyses with Bonferroni correction demonstrated that Reading induced significantly higher Mental Demand than Doodling ($p = .004$), as well as higher Temporal Demand ($p = .013$) and required more Effort ($p = .025$). No significant differences were found for

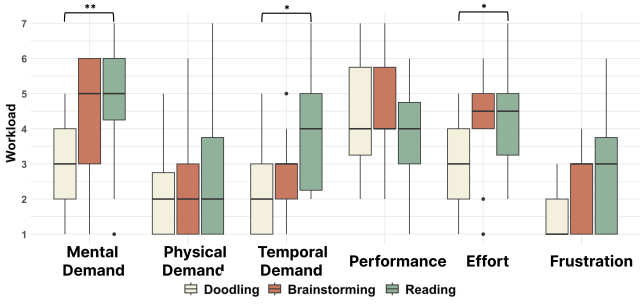

Figure 5: The raw NASA-TLX results ($p < .05$(*), $p < .01$(**).

Table 1: Codebook for Notification Response Patterns in Mixed Reality. The frequency of each code assignment is indicated in parentheses. Individual notifications may receive multiple codes.

Theme	Sub-Theme	Definition
Sender (11)	Authority-Based Prioritisation (8)	Preferential response to notifications from supervisors
	Social Relationship Prioritisation (3)	Preferential response to notifications from friends
	Group Message Ignorance (8)	Tendency to ignore group messages
Content (14)	Action Request Response (12)	Tendency to respond to notifications requiring action or questions
	Content Length Sensitivity (5)	Response patterns influenced by notification length
	Information Density Evaluation (3)	Response based on perceived information value
	Implicit Content Cues (3)	Response influenced by implicit cues of notification
Activity (14)	Cognitive Load Management (4)	Response patterns based on cognitive demands of current activity
	Activity Engagement Level (2)	Response patterns influenced by engagement with current activity
	Activity-Specific Response Strategies (14)	Different response strategies for different MR activities
	Task Disinterest Displacement (3)	Higher response rate due to disinterest in primary task

Physical Demand ($p = .491$). For overall workload, which met normality assumptions, Repeated Measures ANOVA revealed significant differences between activities ($p < .001$). Post-hoc analysis indicated that Brainstorming required significantly higher overall workload than Doodling ($p = .041$), while Reading demanded even more significantly ($p = .004$). No significant difference was found between Reading and Brainstorming ($p = .169$). Our analysis confirms that the three MR activities successfully created varied cognitive demands, with reading imposing the highest workload, followed by Brainstorming and Doodling, validating our experimental design for studying notification behaviour across different cognitive states. These findings support our methodology by confirming we collected data across meaningfully different contexts.

4.6 Implications

Our user study establishes two essential datasets: a self-labelled dataset ($N = 18 \times 90$) and an interaction-based dataset ($N = 18 \times 108$) capturing notification behaviour during three activities with varying cognitive demands. Our NASA-TLX analysis confirms these activities successfully created distinct cognitive workload conditions, with Reading imposing a significantly higher overall workload than Doodling, validating our experimental design and ensuring data collection across meaningfully different cognitive states. This methodological foundation directly supports our subsequent analyses in section 5 by collecting users' self-reported behaviour patterns through the ESM. The confirmed differences in cognitive load across activities will be particularly valuable when analysing how Activity influences notification responsiveness. Furthermore, these datasets provide the necessary training and testing data for developing our personalised notification urgency classifier (PersoNo) in section 6, where we will evaluate different classification approaches using both self-labelled and interaction-based data to determine which yields the most effective urgency predictions.

5 Study 2: Human Behaviour Patterns

While existing research established theoretical frameworks for mobile notifications [25, 41, 34], our work adopts a top-down thematic approach [3] to systematically examine how these frameworks manifest differently in MR. The unique perceptual and contextual factors of MR, where digital information overlays physical space, likely transform how users perceive, prioritise, and respond to notifications. It raises our **RQ1: How do users behave and respond to notifications in MR?** The next contribution is un-

derstanding human behaviour patterns regarding MR notifications. Our novel insights establish the critical contextual variables that must be prioritised when developing MR notification classifiers, directly informing our later prompt construction methodology.

5.1 Thematic Analysis

Two researchers independently reviewed all users' ESM data and developed the initial codebook, referencing previous work by [25, 32]. We identified the Sender, Content, and Activity as the most frequently mentioned factors, which formed the basis of three thematic categories. After independently drafting initial codes, we collaboratively refined definitions and examples through discussion. Subsequently, we conducted a pilot phase on two self-reported behaviour patterns ($\sim$ 10% of the data) to test and refine the codebook.

Following codebook construction, we randomly selected four users' behavioural patterns ($\sim$ 22% within the range suggested by O'Connor et al. [38]) to validate our codebook, ensuring these were distinct from those used in the pilot phase. In addition to the original researchers who developed the codebook, we invited another researcher to code the user behaviour using the established codebook. We calculated Krippendorff's alpha, a standard inter-rater reliability measure for non-mutually exclusive coding schemes [16]. Our Krippendorff's alpha was 0.846 above the standard of reliable labelling results [16]. After validating our codebook's reliability, two researchers independently coded all users' ESM data and subsequently resolved any coding disagreements. Table 1 presents the defined codebook and the frequency of code assignment.

5.2 Results

Our analysis identified three primary themes influencing notification response in MR: Sender, Content, and Activity contexts (see Table 1). Beyond these main themes, participants reported additional notification response patterns, such as opportune timing (N=2). Two participants indicated that opportune timing is crucial in determining whether they would reply to messages within 30 seconds. Interestingly, *P15* mentioned responding to notifications somewhat randomly, even when the content was important, to avoid giving others the impression of constant availability. We categorised these as "Others" and excluded them from our analysis since only a few participants mentioned them.

Participants' likelihood of attending to a notification depended on who the sender was (N=11), but not always in the way seen with smartphones. Surprisingly, personal friendship played a minimal role in the immediate MR notification responses. Only three instances were recorded. This differs markedly from traditional mobile notification research [41, 5], where personal relationships significantly influence notification attendance. Instead, participants gave preferential attention to notifications from their supervisors (N=8), while group notifications were often ignored (N=8).

Content characteristics emerged as equally important, with action request responses (N=12) dominating this theme. Participants also demonstrated sensitivity to content length (N=5) and employed sophisticated information evaluation through implicit cues (N=3) and information density assessment (N=3). While Li et al. [25] found content factors influenced mobile notification preferences more than contextual factors, our results indicate that in MR, content considerations maintain equivalent importance alongside sender and activity contexts, rather than dominating them.

Activity context featured prominently, with activity-specific response strategies (N=14) representing our most frequent code. It indicates that users often developed different notification response rules for various activities in MR. Cognitive load management (N=4) and task disinterest displacement (N=3) revealed how participants balanced attention resources. Two participants reported that they would use the Mute All function in reality when engaging in any activities and respond to the notifications during breaks.

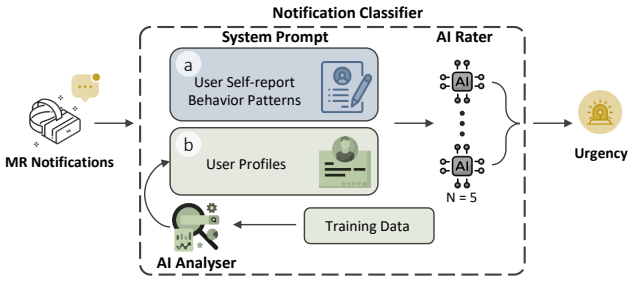

Figure 6: Framework of the Classifiers with different In-Context Learning Prompts. (a) shows $\mathbf{M}_1$ zero-shot learning using user self-reported behaviour patterns. (b) refers to the $\mathbf{M}_2$ multi-agent methods that an AI analyser extracts user profiles from the training data.

5.3 Implications

Our thematic analysis establishes a comprehensive framework that advances MR notification classifier development by identifying key themes (Sender, Content, and Activity) and their critical sub-themes affecting user responsiveness. MR designers can use our codebook to develop more sophisticated classification systems incorporating nuanced factors.

It is noticeable that users' MR notification behaviour patterns differ from their self-reported receptivity to mobile notifications (see subsection 4.3). A substantially higher number of participants considered both Sender and Activity factors when deciding whether to respond to MR notifications. This finding suggests that Activity represents a crucial variable when designing an MR notification classifier, even though Li et al. [25] demonstrated that utilising notification content alone for mobile notification classifiers could achieve reasonable results. The discrepancy highlights the unique contextual considerations necessary for effective notification management in MR compared to traditional mobile settings.

Furthermore, our findings reveal a significant departure from traditional mobile notification patterns. While previous research suggests users typically prioritise responses to friends' messages due to social pressure [5, 41], only a few participants reported their priorities in social relationships. Instead, users overrode this factor with other considerations. Additionally, the substantially higher number of participants who adopted activity-oriented notification response strategies may be attributed to the more disruptive nature of MR notifications, resulting in reduced notification receptivity during different activities. Thus, for the following notification classifier development, in addition to our hypothesised variables (Sender and Content), we also considered the Activity.

6 Study 3: Notification Classifier

We developed an intelligent notification management system with the training dataset (Study 1) and the key contextual factors identified in Study 2. Next, we describe our notification classifiers and evaluate hypotheses related to three essential classifier components: Data (**[H2]**), Context (**[H3]**), and Algorithm (**[H1]**, **[H4]**).

[H1]: Personalised notification classifier will outperform the general notification classifier. **[H2]**: Self-reported urgency ratings will yield classification performance comparable to interaction-based data. **[H3]**: Activity, mentioned as frequently as Content by our participants, will outperform classification performance beyond models using only Sender and Content variables. **[H4]**: LLM analysers will capture user notification response patterns accurately and comprehensively based on our codebook framework.

6.1 Data Preparation

We utilised part of the notification data collected during MR activities as the test data. Following previous work [25], we designated the last six notifications from each activity as testing data

($N = 6 \times 3 = 18$). This methodology emulates real-world scenarios where historical data informs predictive models of human behaviour patterns. For analysis, we prepared three distinct training datasets: **SR**: Self-Report, containing data labelled directly by participants during the self-report phase; $\mathbf{D}_1$ (Dataset 1), an interaction-based notification dataset collected during MR activities with key variables of Sender and Content; and $\mathbf{D}_2$ (Dataset 2), identical to D_1 but incorporating Activity as an additional contextual variable, which subsection 5.3 identified as crucial.

6.2 Model Design

Unlike previous research on notification classifier [7, 33] that implemented traditional Machine Learning (ML) algorithms, we leverage LLMs to build our notification classifier. In our study, we utilised Qwen [58] (QwQ-32B), an open-source LLM, to ensure reproducibility. Pretrained on large-scale corpora, LLMs have demonstrated essential language understanding capabilities [4]. This foundation enables the model to analyse the semantic meaning of notification content and capture general urgency indicators, such as calls to action and time-sensitive content [25]. Furthermore, previous work [4] has shown that In-Context Learning (ICL) enables pre-trained language models to more effectively address downstream tasks, in our case, notification urgency classification. This technique requires a small amount of data, making LLMs ideally suited for personalized notification classification, which inherently has limited training data since all the data are from a single user. Additionally, ICL provides an interpretable interface for LLM interaction [4], which enhances the explainability of the algorithm. Compared with traditional ML approaches, using LLMs with ICL is more aligned with standards of Human-Centered Artificial Intelligence (HCAI) [56], prioritising user understanding and control.

Three models were included in our study. We employed two ICL methods to optimise the LLMs for notification urgency classification: ($\mathbf{M}_1$) Zero-shot learning [4] utilising only user-reported behaviour patterns collected via ESM, and ($\mathbf{M}_2$) Multi Agent (MA) [60] implementing an analyser LLM to create user profiles from training data, which rater LLMs then use to classify test data. Apart from the personalised classifiers using M_1 and M_2, we evaluated general notification classifiers using base models (**Base**) to predict the test dataset directly. These models, trained on general corpora, analyse urgency levels from the general users' perspective.

To further improve the classification reliability and accuracy, we integrated homogeneous LLM self-ensemble techniques [60] (see Figure 6) and Chain-of-Thoughts (CoT) prompting [62]. Each method employed five raters with a temperature setting of 1, resulting in five independent votes per notification with certain randomness. It was specifically employed to reduce prediction variance across multiple raters. The final urgency label was determined by majority vote. Rather than direct classification, CoT prompted LLMs to articulate their reasoning process step-by-step before making predictions, significantly improving transparency and accuracy.

6.3 Prompt Design

We designed two ICL prompts as the base prompts based on D_1/SR and D_2. We augmented these base prompts with specific information to construct complete prompts for each methodological test, such as "The following is the user behaviour pattern: {user_pattern}." While being used for the Base, we left the {user_pattern} blank. The prompts hold the variables in different datasets: the first prompt (**P1**) instructed the model to analyse only the *Sender* and *Content* variables, suitable for SR and D_1, while the second prompt (**P2**) expanded the analysis to include *Sender*, *Content*, and *Activity* variables, corresponding to D_2. These prompts guided models in formulating reasoning based on the respective variables before generating predictions. Further, for the analyser LLMs in the MA framework, we integrated findings from the thematic analysis (see section 5), instructing the analyser to consider all identified sub-themes when developing comprehensive user profiles. We applied a similar method to design the rater prompts.

Table 2: Comparison of models

	Base		**M_1**		**M_2**		
	D_1	D_2	D_1	D_2	SR	D_1	D_2
Accuracy	0.670	0.670	0.735	0.679	0.747	0.759	**0.815**
FNR	0.703	0.627	0.528	0.614	0.432	0.550	**0.381**
Specificity	0.887	0.846	0.838	0.818	0.821	**0.923**	0.875
AUROC	0.595	0.609	0.684	0.617	0.721	0.708	**0.786**

6.4 Model Configurations

We prompted models using factorial combinations of methods and datasets METHOD × DATASET. For example, with $M_2 \times D_1$, we employed the multi-agent (M_2) method and prompted analyser models with interaction-based data containing Sender and Content variables (D_1) using P1-based prompts. More specifically, M_1 operates without a training notification dataset, relying exclusively on user-reported behavioural patterns. Consequently, we utilised both prompts (P1 and P2, detailed in subsection 6.3) to evaluate this method, with P1 analysing Sender and Content variables while P2 incorporated Activity as an additional contextual factor. Additionally, we evaluated a general notification classifier (Base) that, similar to M_1, employed both P1 and P2 prompts but without any information about user profiles. These Base models relied solely on their pre-training on general corpora to analyse notification urgency from a non-personalised perspective.

6.5 Results

We evaluated model performance using Accuracy, False Negative Rate (FNR), Specificity, and Area Under the Receiver Operating Characteristic (AUROC). While accuracy measures overall correctness, it can be misleading with imbalanced notification urgency classes. Specificity (true negative rate) measures the system's ability to filter out non-urgent notifications, directly addressing our goal of reducing unnecessary interruptions in MR. Furthermore, we incorporated the FNR to quantify missed urgent notifications. Based on previous work's findings [20, 34] that users tolerated interruptions to avoid missing important updates, we assumed that the cost of missing important notifications was higher than the cost of being disturbed and prioritised minimising FNR. AUROC (see Figure 8) provides a threshold-independent assessment of discriminative capability, remaining robust against the common imbalance where non-urgent notifications typically outnumber urgent ones.

See Table 2 for the performance of all model configurations. To evaluate our hypotheses regarding Data (**[H2]**) and Context (**[H3]**), we primarily focused on comparing *PersoNo* (M_2), our proposed notification system, which demonstrated superior performance across all metrics compared to alternative models. Statistical significance between conditions was assessed using pairwise t-tests.

For **[H1]**, we compared personalised notification classifiers against the general classifier (Base). Results showed that personalised approaches (M_1 and M_2) consistently outperformed the base model across all metrics. The base model achieved only 0.670 accuracy with P1 and P2, while personalised models reached up to 0.815 accuracy with M_2 on D_2. Notably, the FNR of the base model was substantially higher (0.586 and 0.523) than personalised approaches, indicating that users would be more likely to miss important information using the general classifier.

For **[H2]**, we examined whether self-reported urgency ratings yield comparable performance to interaction-based data. The M_2 model using SR achieved 0.747 accuracy and 0.721 AUROC, while the interaction-based data (D_2) resulted in 0.759 accuracy and 0.708 AUROC. Pairwise t-tests showed no significant differences between these approaches in accuracy and FNR ($t(17) = .334$, $p = .742$ and

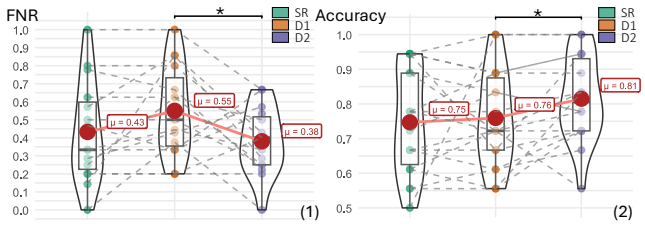

Figure 7: Performance comparison of M_2 (*PersoNo*). (1) False Negative Rate and (2) Accuracy across conditions. Dashed lines indicate changes in individual participant results ($p < .05$(*)).

$t(17) = 1.99$, $p = .066$), supporting our hypothesis that self-reports serve as viable alternatives for classifiers.

For **[H3]**, we investigated whether incorporating Activity context would improve classification performance. The model trained on D_2 (which included Activity context) significantly outperformed D_1 (Sender and Content only) in both FNR (0.381 compared to 0.550, $t(17) = 2.30$, $p = .037$) and accuracy (0.815 compared to 0.759, $t(17) = -2.15$, $p = .046$). Moreover, D_2 demonstrated enhanced AUROC (0.786 compared to 0.708), confirming that Activity represents a crucial contextual factor in MR notification management. Figure 7 (1) demonstrates a reduction in FNR for the majority of participants, as indicated by the general downward trend of the gray dashed lines, while (2) exhibits the opposite pattern, revealing an improvement in Accuracy.

For **[H4]**, we assessed whether LLM analysers could effectively capture user notification response patterns. The superior performance of M_2, consisting of LLM analysers generating user profiles and raters providing predictions, validated this hypothesis. M_2 consistently outperformed M_1 prompted by P2, with the D_2 configuration achieving significantly higher accuracy (0.815, $t(17) = 3.335$, $p = .004$) and lower FNR (0.381, $t(17) = -3.008$, $p = .009$).

6.6 Implications

Analysis of our experimental results provides substantial evidence supporting all four hypotheses. However, we found some unexpected decreases in the model performance with the incorporation of the Activity context. Table 2 shows the accuracy of M_1 decreased from 0.735 to 0.679 when LLM raters were prompted to consider Activity context, which appears to contradict our **[H3]**.

Several factors can explain this apparent contradiction. First, the ESM data revealed that not all participants reported activity-based behavioural patterns regarding MR notifications. In this context, incorporating the Activity component would likely complicate the rating process and potentially bias final evaluations by introducing redundant information. Second, participants' self-reported activity-oriented behavioural patterns were often expressed in broad terms that lacked the specificity needed for fine-grained classification. This generality made it difficult for LLM raters to detect subtle differences between individual users' notification preferences across various activities. When LLMs were instructed to consider Activity without sufficient user-specific data, they likely defaulted to general population assumptions rather than personalised predictions. For instance, LLMs might automatically classify notifications during Reading as non-urgent based on its typical high cognitive load, regardless of individual user preferences. However, some participants reported difficulty focusing on the Reading task due to a lack of interest, which increased their willingness to respond to IMs. Furthermore, the zero-shot learning approach employed in M_1 appears particularly sensitive to this issue, as it lacks training examples that would help calibrate activity-based predictions to individual users.

Nevertheless, the more sophisticated M_2 approach, which incorporated explicit training examples, successfully leveraged Activity context to improve classification performance, ultimately supporting **[H3]** within more structured learning frameworks.

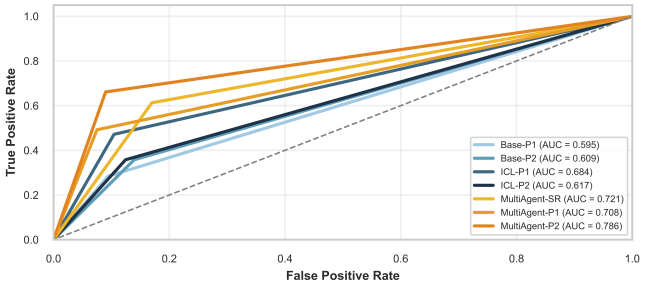

Figure 8: AUROC Results. TPR represents the True Positive Rate, and FPR refers to the False Positive Rate.

7 Discussion

7.1 Human-Centered PersoNo

The *PersoNo* system, comprising an LLM analyser and raters, demonstrated superior performance (accuracy of 0.815 and FNR of 0.381 with M_2-D_2), highlighting the potential of LLM applications in personalised systems. Furthermore, the interpretability of LLM-based systems represents a significant advantage. As stated by Shneiderman [56], HCAI should not only focus on the algorithmic performance but also empower users by offering control and understanding. Unlike black-box ML models, our CoT approach made the classification process transparent, with LLM raters explicitly articulating their reasoning before making predictions. This transparency improved classification accuracy and could also enhance user trust and system adoption. Additionally, our multi-agent architecture enhances the system's human-centered qualities by separating profile generation from notification classification. This separation enables potential user intervention in profile creation, allowing users to review and adjust automatically extracted preferences. By aligning fundamental aspects of HCAI design, such capability affords users direct control over how the system interprets their notification behaviour, fostering transparency and autonomy.

7.2 Impacts and Applications of PersoNo

Advancing Notifications Management from Mobile to MR Our work extends mobile intelligent notification systems research [25, 32] to the unique challenges of MR environments. While prior mobile notification classifiers such as PrefMiner [32] and content-driven systems [33] established foundational approaches for managing interruptions, they operated under fundamentally different interaction paradigms. Mobile users can physically distance themselves from devices, but MR notifications directly overlay the user's visual field. PersoNo addresses this gap by introducing the personalised notification classifier specifically designed for MR's spatial computing context. Our finding that activity context equals content importance (see section 5) gives different perspectives from Li et al.'s [25] well-known mobile-centric framework, where content dominated other factors. This shift reflects the fundamental difference in how notifications compete for cognitive resources in MR, a finding that extends Lindlbauer et al.'s [28] work on context-aware MR interfaces to the notification domain.

Redefining Personalisation Via Limited Data Learning Traditional personalised notification systems required extensive data collection periods. For example, Mehrotra et al. [32] needed 15 days, while Pielot et al. [40] averaged four weeks. This requirement has been a significant barrier to adoption, as noted by Maister [30]. PersoNo fundamentally reconceptualises the notification classification through LLM-based learning, achieving 81.5% accuracy with just 90 training instances (**[H4]**). It demonstrates the superior performance with a limited dataset. This advancement builds upon

recent work in few-shot learning [4] but applies it specifically to the HCI challenge of notification management. Our multi-agent architecture (M_2) represents a novel application of LLM capabilities to extract meaningful user profiles from minimal data—a contribution that extends beyond notification systems to any personalised intelligent interface requiring rapid adaptation to individual users.

Bridging VR/MR Notification Design and MR Intelligence While VR and MR notification research has advanced placement strategies [18, 52] and multimodal design [13], these studies primarily addressed notification presentation rather than intelligent filtering. PersoNo bridges this critical gap by introducing personalised urgency classification, which enables dynamic adaptation of established notification design principles. Our system transforms static design guidelines into context-aware behaviours: urgent notifications leverage the bottom-center placement proven most noticeable [53, 45], while non-urgent messages can utilise less intrusive in-situ positioning [52] that preserves spatial context without demanding immediate attention. In addition to the adaptive notification placement, the future VR/MR developers may consider pushing the non-urgent notifications until the opportune time, such as the break during VR/MR activities [7].

This urgency-based adaptation directly supports the vision of calm technology in MR [23], where information should inform without overwhelming. By filtering notifications before they reach the presentation layer, PersoNo ensures that only contextually appropriate interruptions utilise prime visual real estate. Furthermore, the system's classification output can further integrate with advanced MR adaptation frameworks like SituationAdapt [27], which employs vision-and-language models for environmental analysis. While PersoNo determines notification urgency based on user patterns, SituationAdapt could identify optimal spatial placement by analysing the user's current visual scene, creating a comprehensive pipeline from urgency assessment to context-aware positioning.

Our current implementation employs binary urgency classification, aligning with established notification research methodologies [33]. However, the modular architecture of our approach facilitates future extensions to multi-level urgency schemes through prompt-based redefinition of urgency categories. Such granular classification would unlock the full potential of existing VR/MR notification placement [52], enabling nuanced placement strategies where notification position, opacity, and persistence vary along an urgency continuum rather than a simple binary threshold.

7.3 Subjective Labelling and Objective User Behaviour

Our study compared two data collection approaches for training notification classifiers: self-labelled urgency ratings (subjective) and actual interaction behaviour (objective). Although both datasets were employed in previous notification systems [7, 40], prior subjective data were collected primarily through ESM. Consequently, the literature lacks analysis of user-labelled datasets and their comparison with actual interaction behaviour datasets. The comparable performance between models trained on self-reported data (M_2-SR) and interaction-based data (M_2-D_1) suggests that self-reporting can effectively substitute for longer-term interaction tracking when building personalised notification classifiers (**[H2]**). This finding contradicts prior work suggesting weak correlations between self-reporting and actual behaviours [10]. The unexpected consistency between subjective and objective data may be due to our study's focus on the specific interaction behaviour of notification responses. Users are highly familiar with notification patterns in their everyday messaging interactions and are likely to possess a strong awareness of both their notification preferences and behavioural tendencies.

Thus, rather than extended data collection periods spanning weeks, which has been standard practice in previous notification research [32, 35], personalised notification systems could be deployed using simple user-provided labels. In addition to PersoNo's ability to classify notification urgency accurately with limited personal data, this approach could significantly improve user acceptance of intelligent notification systems by minimising the waiting period before deployment. Beyond the notification, a potential implication is that all intelligent systems could leverage subjective feedback as training data for interactions with users who are highly familiar with them. We acknowledge that human behaviour patterns vary over time [32, 33]. However, we propose using self-labelled data to initiate and facilitate user adoption of the personalised system. Such implementations can continuously adapt based on users' interaction data to accommodate changing behaviour patterns.

7.4 Limitations

Our study establishes foundational insights for MR notification management while revealing opportunities for future research. First, our self-labelling methodology focused on content and sender variables, following established mobile notification research paradigms. While this approach successfully demonstrated the viability of self-reported data for classifier training, future work could enhance this methodology by incorporating activity-specific rating scenarios to capture the full contextual richness we identified as crucial for MR environments. Second, our controlled laboratory environment and pre-established social relationships (using participant-provided names as placeholders) ensured systematic variable manipulation but may not fully capture real-world MR notification dynamics. In-the-wild deployments with a longitudinal study would provide valuable insights into PersoNo's performance under authentic conditions and evolving social relationships. Finally, while our study identified three key contextual dimensions (content, sender, activity) with 18 participants, notification receptivity in MR likely depends on additional factors. Future work could explore broader contextual variables, including temporal factors (time of day) [54], spatial contexts (location [33], proximity to others), and MR-specific factors (immersion level, virtual-physical task integration), while validating findings with larger sample sizes.

8 Conclusion

Our research addresses notification management in MR by introducing *PersoNo*, a personalised LLM-based notification urgency classifier. Through user studies, we collected the first MR notification dataset and discovered that activity context is equally important as content and sender in MR, which is a key difference from mobile notification management. *PersoNo* achieved 81.5% accuracy, 0.786 AUROC, and 0.381 FNR by effectively analysing user profiles from limited data, outperforming baseline approaches. Notably, with *PersoNo*, self-reported urgency ratings proved effective for classifier training while considering the contexts, contradicting assumptions about self-report validity. To conclude, adhering to HCAI design principles, *PersoNo* employs AI to minimise distractions while ensuring notification awareness, simultaneously providing users with understanding and control.

Acknowledgments

This research was supported by the Hong Kong Polytechnic University's Start-up Fund for New Recruits (No. P0046056), Departmental General Research Fund (DGRF) from HK PolyU ISE (No. P0056354), and PolyU RIAM – Research Institute for Advanced Manufacturing (No. P0056767). Jingyao Zheng and Xian Wang were supported by a grant from the PolyU Research Committee under student account codes RMCU and RMHD, respectively. This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

REFERENCES

[1] J. Alonzo, G. Tindal, K. Ulmer, and A. Glasgow. easycbm online progress monitoring assessment system. eugene, or: Center for educational assessment accountability, 2006. 4

[2] S. Aminikhanghahi, R. Fallahzadeh, M. Sawyer, D. J. Cook, and L. B. Holder. Thyme: Improving smartphone prompt timing through activity awareness. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 315–322. IEEE, 2017. doi: 10.1109/ICMLA.2017.0-141 2

[3] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006. doi: 10.1191/1478088706qp063oa 5

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 7, 9

[5] X.-J. Chang, F.-H. Hsu, E.-C. Liang, Z.-Y. Chiou, H.-H. Chuang, F.-C. Tseng, Y.-H. Lin, and Y.-J. Chang. Not merely deemed as distraction: Investigating smartphone users' motivations for notification-interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2023. doi: 10.1145/3544548.3581146 2, 4, 6

[6] Y.-J. Chang, Y.-J. Chung, and Y.-H. Shih. I think it's her: Investigating smartphone users' speculation about phone notifications and its influence on attendance. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–13, 2019. doi: 10.1145/3338286.3340125 1, 2, 3

[7] K.-W. Chen, Y.-J. Chang, and L. Chan. Predicting opportune moments to deliver notifications in virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, New York, NY, USA, 2022. doi: 10.1145/3491102.3517529 1, 2, 3, 4, 7, 9

[8] X. Chen, N. Srivastava, R. Jain, J. Healey, and T. Dingler. Characteristics of deep and skim reading on smartphones vs. desktop: A comparative study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. ACM, New York, NY, USA, 2023. doi: 10.1145/3544548.3581174 4

[9] H. Cho, D. Edgar, D. Lindlbauer, and J. O'Hagan. Evaluating dynamic delivery of audio+ visual message notifications in xr. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 277–287. IEEE, 2025. doi: 10.1109/VR59515.2025.00052 2

[10] J. Dang, K. M. King, and M. Inzlicht. Why are self-report and behavioral measures weakly correlated? *Trends in cognitive sciences*, 24(4):267–269, 2020. doi: 10.1016/j.tics.2020.01.007 2, 9

[11] M. Dredze, T. Brooks, J. Carroll, J. Magarick, J. Blitzer, and F. Pereira. Intelligent email: Reply and attachment prediction. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pp. 321–324, 2008. doi: 10.1145/1378773.1378820 1, 2

[12] C. George, P. Tamunjoh, and H. Hussmann. Invisible boundaries for vr: Auditory and haptic signals as indicators for real world boundaries. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3414–3422, 2020. doi: 10.1109/TVCG.2020.3023607 2

[13] S. Ghosh, L. Winston, N. Panchal, P. Kimura-Thollander, J. Hotnog, D. Cheong, G. Reyes, and G. D. Abowd. Notifivr: Exploring interruptions and notifications in virtual reality. *IEEE transactions on visualization and computer graphics*, 24(4):1447–1456, 2018. doi: 10.1109/TVCG.2018.2793698 1, 2, 9

[14] M. Gonzalez-Franco, R. Pizarro, J. Cermeron, K. Li, J. Thorn, W. Hutabarat, A. Tiwari, and P. Bermell-Garcia. Immersive mixed reality for manufacturing training. *Frontiers in Robotics and AI*, 4:3, 2017. doi: 10.3389/frobt.2017.00003 1

[15] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, vol. 52, pp. 139–183. Elsevier, 1988. doi: 10.1016/S0166-4115(08)62386-9 5

[16] A. F. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007. doi: 10.1080/19312450709336664 6

[17] J. M. Hektner, J. A. Schmidt, and M. Csikszentmihalyi. *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007. 3, 5

[18] C.-Y. Hsieh, Y.-S. Chiang, H.-Y. Chiu, and Y.-J. Chang. Bridging the virtual and real worlds: A preliminary study of messaging notifications in virtual reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–14. ACM, New York, NY, USA, 2020. doi: 10.1145/3313831.3376228 1, 2, 9

[19] S. Imamov, D. Monzel, and W. S. Lages. Where to display? how interface position affects comfort and task switching time on glanceable interfaces. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 851–858. IEEE, 2020. doi: 10.1109/VR46266.2020.00110 2

[20] S. T. Iqbal and E. Horvitz. Notifications and awareness: a field study of alert usage and preferences. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pp. 27–30, 2010. doi: 10.1145/1718918.1718926 7

[21] N. Joshi and D. Vogel. Constrained highlighting in a document reader can improve reading comprehension. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2024. doi: 10.1145/3613904.3642314 4

[22] A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, et al. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 955–964, 2016. doi: 10.1145/2939672.2939801 4

[23] K. Y. Lam, L. H. Lee, and P. Hui. A2w: Context-aware recommendation system for mobile augmented reality web browser. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, p. 2447–2455. ACM, New York, NY, USA, 2021. doi: 10.1145/3474085.3475413 9

[24] K. Lee, H. Li, M. R. Wellyanto, Y. J. Tham, A. Monroy-Hernández, F. Liu, B. A. Smith, and R. Vaish. Exploring immersive interpersonal communication via ar. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–25, 2023. doi: 10.1145/3579483 3

[25] T. Li, J. K. Haines, M. F. R. De Eguino, J. I. Hong, and J. Nichols. Alert now or never: Understanding and predicting notification preferences of smartphone users. *ACM Transactions on Computer-Human Interaction*, 29(5):1–33, 2023. doi: 10.1145/3478868 1, 2, 3, 4, 5, 6, 7, 8

[26] Z. Li, Y. F. Cheng, Y. Yan, and D. Lindlbauer. Predicting the noticeability of dynamic virtual elements in virtual reality. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2024. doi: 10.1145/3613904.3642399 2

[27] Z. Li, C. Gebhardt, Y. Inglin, N. Steck, P. Streli, and C. Holz. Situationadapt: Contextual ui optimization in mixed reality with situation awareness via llm reasoning. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. ACM, New York, NY, USA, 2024. doi: 10.1145/3654777.3676470 9

[28] D. Lindlbauer, A. M. Feit, and O. Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pp. 147–160, 2019. doi: 10.1145/3332165.3347945 1, 2, 4, 5, 8

[29] M. J. Maas and J. M. Hughes. Virtual, augmented and mixed reality in k–12 education: A review of the literature. *Technology, Pedagogy and Education*, 29(2):231–249, 2020. doi: 10.1080/1475939X.2020.1737210 1

[30] D. H. Maister et al. *The psychology of waiting lines*. Harvard Business School Boston, 1984. 2, 8

[31] B. Marques, S. Silva, P. Dias, and B. Sousa-Santos. Which notification is better? comparing visual, audio and tactile cues for asynchronous mixed reality (mr) remote collaboration: A user study. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*, MUM '22, p. 276–278. ACM, New York, NY, USA, 2022. doi: 10.1145/3568444.3570587 1, 2

[32] A. Mehrotra, R. Hendley, and M. Musolesi. PrefMiner: mining user's preferences for intelligent mobile notification management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 1223–1234. ACM, Heidelberg Germany, September 2016. doi: 10.1145/2971648.2971747 2, 6, 8, 9

[33] A. Mehrotra, M. Musolesi, R. Hendley, and V. Pejovic. Designing

content-driven intelligent notification mechanisms for mobile applications. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 813–824, 2015. doi: 10.1145/2750858.2807544 1, 2, 3, 7, 8, 9

[34] A. Mehrotra, V. Pejovic, J. Vermeulen, R. Hendley, and M. Musolesi. My phone and me: Understanding people's receptivity to mobile notifications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, p. 1021–1032. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858566 2, 3, 4, 5, 7

[35] H. Oh, L. Jalali, and R. Jain. An intelligent notification system using context from real-time personal activity monitoring. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2015. 1, 9

[36] T. Okoshi, K. Tsubouchi, M. Taji, T. Ichikawa, and H. Tokuda. Attention and engagement-awareness in the wild: A large-scale study with adaptive notifications. In *2017 ieee international conference on pervasive computing and communications (percom)*, pp. 100–110. IEEE, 2017. doi: 10.1109/PERCOM.2017.7917856 2

[37] J. Orlosky, K. Kiyokawa, and H. Takemura. Managing mobile text in head mounted displays: studies on visual preference and text placement. *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(2):20–31, 2014. doi: 10.1145/2636242.2636246 2

[38] C. O'Connor and H. Joffe. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods*, 19:1609406919899220, 2020. 6

[39] V. Pejovic and M. Musolesi. Interruptme: designing intelligent prompting mechanisms for pervasive applications. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 897–908, 2014. doi: 10.1145/2632048.2632062 2

[40] M. Pielot, B. Cardoso, K. Katevas, J. Serrà, A. Matic, and N. Oliver. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–25, 2017. doi: 10.1145/3130956 2, 8, 9

[41] M. Pielot, K. Church, and R. de Oliveira. An in-situ study of mobile phone notifications. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*, MobileHCI '14, p. 233–242. ACM, New York, NY, USA, 2014. doi: 10.1145/2628363.2628364 2, 5, 6

[42] M. Pielot and L. Rello. Productive, anxious, lonely: 24 hours without push notifications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pp. 1–11, 2017. doi: 10.1145/3098279.3098526 1, 2

[43] M. Pielot, A. Vradi, and S. Park. Dismissed! a detailed exploration of how mobile phone users handle push notifications. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '18. ACM, New York, NY, USA, 2018. doi: 10.1145/3229434.3229445 3

[44] L. Plabst, F. Niebling, S. Oberdörfer, and F. Ortega. Order up! multimodal interaction techniques for notifications in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 31(5):2258–2267, 2025. doi: 10.1109/TVCG.2025.3549186 1

[45] L. Plabst, S. Oberdörfer, F. R. Ortega, and F. Niebling. Push the red button: Comparing notification placement with augmented and non-augmented tasks in ar. In *Proceedings of the 2022 ACM Symposium on Spatial User Interaction*, SUI '22. ACM, New York, NY, USA, 2022. doi: 10.1145/3565970.3567701 4, 9

[46] B. Poppinga, W. Heuten, and S. Boll. Sensor-based identification of opportune moments for triggering notifications. *IEEE Pervasive Computing*, 13(1):22–29, 2014. doi: 10.1109/MPRV.2014.15 2

[47] S. Pradhan, L. Qiu, A. Parate, and K.-H. Kim. Understanding and managing notifications. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9. IEEE, New York, NY, USA, 2017. doi: 10.1109/INFOCOM.2017.8057231 2

[48] Z. Qu, R. Byrne, and M. Gorlatova. "looking" into attention patterns in extended reality: An eye tracking-based study. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 855–864, Oct 2024. doi: 10.1109/ISMAR62088.2024.00101 3

[49] K. Quinn and J. L. Gabbard. Augmented reality visualization techniques for attention guidance to out-of-view objects: A systematic review. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 826–835, Oct 2024. doi: 10.1109/ISMAR62088.2024.00098 3

[50] E. F. Rietzschel, B. A. Nijstad, and W. Stroebe. Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental social psychology*, 42(2):244–251, 2006. doi: 10.1016/j.jesp.2005.04.005 4

[51] R. Rzayev, S. Korbely, M. Maul, A. Schark, V. Schwind, and N. Henze. Effects of position and alignment of notifications on ar glasses during social interaction. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, NordiCHI '20. ACM, New York, NY, USA, 2020. doi: 10.1145/3419249.3420095 1, 2

[52] R. Rzayev, S. Mayer, C. Krauter, and N. Henze. Notification in vr: The effect of notification placement, task and environment. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, CHI PLAY '19, p. 199–211. ACM, New York, NY, USA, 2019. doi: 10.1145/3311350.3347190 1, 2, 3, 4, 5, 9

[53] R. Rzayev, P. W. Woźniak, T. Dingler, and N. Henze. Reading on smart glasses: The effect of text position, presentation type and walking. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–9, 2018. doi: 10.1145/3173574.3173619 4, 9

[54] H. Sarker, M. Sharmin, A. A. Ali, M. M. Rahman, R. Bari, S. M. Hossain, and S. Kumar. Assessing the availability of users to engage in just-in-time intervention in the natural environment. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pp. 909–920, 2014. doi: 10.1145/2632048.2636082 2, 9

[55] I. H. Sarker, M. A. Kabir, A. Colman, and J. Han. Designing architecture of a rule-based system for managing phone call interruptions. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, p. 898–903. ACM, New York, NY, USA, 2017. doi: 10.1145/3123024.3124562 2

[56] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020. doi: 10.1080/10447318.2020.1741118 7, 8

[57] M. Speicher, B. D. Hall, and M. Nebeling. What is mixed reality? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–15. ACM, New York, NY, USA, 2019. doi: 10.1145/3290605.3300767 1

[58] Q. Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025. 7

[59] K. Vertanen and P. O. Kristensson. Mining, analyzing, and modeling text written on mobile devices. *Natural Language Engineering*, 27(1):1–33, 2021. doi: 10.1017/S1351324919000548 3

[60] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. 7

[61] D. Weber, A. Voit, G. Kollotzek, and N. Henze. Annotif: A system for annotating mobile notifcations in user studies. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, pp. 1–12, 2019. doi: 10.1145/3365610.3365611 3

[62] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 7

[63] R. Wu and H.-T. Chen. The effect of visual and auditory modality mismatching between distraction and warning on pedestrian street crossing behavior. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 1045–1054, Oct 2023. doi: 10.1109/ISMAR59233.2023.00121 3