Granular Feedback in Human–AI Interaction: Exploring Trade-offs in Ambient Intelligence

Kevin Delcourt
DIRO
Université de Montréal
Montréal, QC, Canada
kevin.delcourt@umontreal.ca

Mathieu Renard
Centre Génie Industriel
IMT Mines Albi
Albi, France
mathieu.renard@mines-albi.fr

Sylvie Trouilhet
IRIT
Université de Toulouse
Toulouse, France
Sylvie.Trouilhet@irit.fr

Jean-Paul Arcangeli
IRIT
Université de Toulouse
Toulouse, France
Jean-Paul.Arcangeli@irit.fr

***Abstract*—Ambient Intelligence (AmI) systems are ubiquitous environments that sense, adapt to, and respond to users' needs. Although these systems increasingly rely on human feedback for machine learning and personalization, determining how and what feedback to elicit remains a major challenge. While fine-grained feedback is assumed to improve machine learning performance by providing richer learning signals, its impact on user experience remains underexplored. We investigate how feedback granularity affects both usability and machine learning ability of an adaptive system for opportunistic software composition: a novel AmI approach that dynamically delivers tailored services by combining available components to meet implicit user goals.**

We conducted two complementary empirical studies in a simulated ambient environment, comparing a baseline interaction with a condition that allows users to provide feedback at different levels of granularity on the services delivered by the intelligent system. The first study evaluates user experience using the USE questionnaire, while the second analyzes the AmI system's learning behavior. Our results indicate that introducing granular feedback tends to decrease perceived usability compared to the baseline, while simultaneously providing benefits to system learning performance. These findings suggest that, in human-in-the-loop AmI systems and more generally in human-AI systems, increasing user control does not necessarily lead to better overall interaction quality. This highlight a clear usability–performance trade-off. We discuss design implications for calibrating feedback mechanisms in adaptive Human–AI systems, emphasizing the need to balance learning benefits against usability costs in everyday, non-expert contexts.

***Index Terms*—Human-in-the-loop, Usability, User Studies, Reinforcement Learning, Recommender Systems.**

I. Introduction

Recent work in human–AI collaboration highlights a trend toward proactive AI agents that go beyond simple reactive responses—anticipating user needs, leveraging contextual information, and initiating actions based on learned models of user goals—marking a significant shift in how AI systems interact with users in dynamic environments [1]. For such systems, the quality of decisions and actions depends critically on the quality of their interactions with humans. As a result, applications that integrate humans in the loop increasingly rely on user feedback to guide learning and decision-making [2]. In many cases, however, this feedback is very simple, such as a binary "good" or "bad" signal [3].

Granular feedback, by contrast, refers to fine-grained input that targets specific aspects of an AI system's output rather than evaluating it as a whole, thereby conveying more detailed information [4]. For example, in an AI programming assistant that recommends code snippets [5], granular feedback may apply to an individual line of code, whereas coarse feedback concerns the entire snippet. Providing mechanisms for granular feedback is recommended in established guidelines for the design of usable human–AI systems [6].

Granular feedback is particularly relevant in the context of **Ambient Intelligence** (AmI) [7], which extends human–AI interaction to pervasive, adaptive environments such as modern healthcare environments [8], smart homes [9], or smart campuses [10]. In these environments, designing usable and effective interactions is a central challenge [11], [12], and incorporating granular feedback has been suggested as a promising way to improve system adaptability and user experience [13]. At the same time, increasing feedback granularity may introduce interaction problems: while it can improve system performance, it may also clutter the interface and reduce usability [14]. However, very few studies provide controlled, empirical comparisons of feedback granularity in terms of both machine learning performance and usability metrics. This limits our understanding of how the design of feedback mechanisms mediates the trade-off between task effectiveness and human factors, leaving open questions about optimal feedback granularity in human-AI and AmI applications.

In this paper, we investigate these questions in the context of opportunistic software composition [13], a novel approach for dynamically constructing applications on-the-fly in ambient environments. Using an opportunistic software composition prototype, the **Opportunistic Composition Engine** (OCE) [15] as a testbed, we study the effect of incorporating a granular feedback mechanism on both system performance and user experience. We conduct a series of user studies in a simulated ambient environment, comparing a baseline OCE system with a version that allows users to provide granular feedback. Our goal is thus to better understand the relationship between feedback granularity, usability, and learning performance. Our results show that granular feedback improves the richness

of system learning signals but comes at a measurable cost to usability, highlighting the inherent trade-offs in designing human-in-the-loop AmI systems.

We make the following contributions:

- An empirical usability study ($N_1 = 61$) comparing a baseline interaction with a granular feedback condition in OCE, showing that increased feedback expressiveness can negatively affect perceived usability.
- An complementary evaluation of learning performance ($N_2 = 12$), analyzing how granular feedback influences the recommendation system's learning behavior, indicating a potential improvement in learning signals despite reduced usability.
- A quantified trade-off between usability and learning performance in a human-in-the-loop AmI system, highlighting a misalignment between system-centric optimization and human-centered interaction quality.
- Design implications for Human–AmI interaction, suggesting that lightweight or implicit feedback mechanisms may offer a better balance than fine-grained user control in non-expert, everyday contexts.

In the context of Ambient Intelligence, systems are expected to operate seamlessly within users' environments while remaining adaptive, context-aware, and minimally intrusive. In such settings, evaluating the balance between system-driven intelligence (e.g., automation accuracy) and human-centered factors (e.g., usability and user control) is critical. Poorly calibrated systems may either overwhelm users with interaction demands or, conversely, reduce transparency and trust through excessive automation. This work contributes to a deeper understanding of how opportunistic and compositional systems can align with AmI principles. Specifically, it informs the design of systems that can dynamically adapt to user needs without imposing continuous interaction burdens, thereby supporting the broader goal of creating intelligent environments that are both effective and unobtrusive.

II. Related Work

Human-AI approaches increasingly leverage user feedback to guide AI learning and improve task performance. For example, Yu et al. [16] compared scalar, detailed feedback with binary, coarse feedback in reinforcement learning, demonstrating that richer, multi-level feedback can enhance learning performance when appropriately modeled. Looking at the implications of soliciting too much feedback, Honeycutt et al. [17] found that while it can improve model outcomes, soliciting user feedback may negatively affect user trust and perceived accuracy, highlighting a potential trade-off between system performance and subjective usability. Research on interface design in human-AI systems further suggests that increasing the granularity of feedback can impose cognitive load and reduce usability if not carefully managed [14], [18].

In the context of Ambient Intelligence, systems learn from sensor data to personalize services in highly complex and dynamic environments. To adapt rapidly to users' evolving needs, leveraging human feedback is essential. Karami et al. [11] demonstrate that integrating user feedback in smart homes improves adaptation quality, but designing feedback channels that are both effective and unobtrusive remains a challenge. Alves Lino et al. [19] and Sadri [7] emphasize that complex or highly adaptive ambient systems can overwhelm users if not carefully designed, highlighting the need to consider usability in interaction design. Furthermore, Park and Han [12] found that overly complex adaptive behaviors can degrade usability, suggesting that balancing system responsiveness with human cognitive load is a core design challenge in AmI.

Despite this growing body of work in human–AI and AmI interaction design, only a few of these studies empirically compare feedback granularity in terms of learning performance and usability, leaving open questions about its optimal design and motivating our research.

III. Problem

Although prior work highlights the potential benefits of human feedback in adaptive systems, the impact of varying feedback granularity on both learning performance and usability remains underexplored.

The central question is not whether granular feedback can improve learning outcomes—this has been demonstrated in various human-AI contexts [5], [16]—but rather whether users are willing and able to provide such detailed feedback without compromising the overall quality of the interaction. In other words, we are interested in understanding the trade-off between enabling granular user input and maintaining usability in adaptive environments. Based on this problem framing, we formulate the following research questions:

> **RQ1 - Usability.** How does the introduction of granular feedback options affect users' perceived usability and interaction experience in an adaptive Ambient Intelligence system?
> **RQ2 - Learning Performance.** How does granular user feedback influence the learning behavior and performance of the AmI system's recommendation mechanism?
> **RQ3 - Trade-off.** What trade-offs emerge between usability and learning performance when increasing feedback granularity in Human–AmI interaction?

IV. Opportunistic Software Composition and OCE

To answer these research questions, we focus on opportunistic software composition, an Ambient Intelligence approach that dynamically constructs services by combining available software components in the environment. This approach is implemented in the Opportunistic Composition Engine (OCE) [15], which has been demonstrated in realistic AmI scenarios such as opportunistic car navigation [20].

A. Key Principles

In ambient environments, software components (e.g., IoT devices) are often present but not pre-integrated. Opportunistic software composition aims to assemble these components into functioning applications tailored to the user's needs, rather

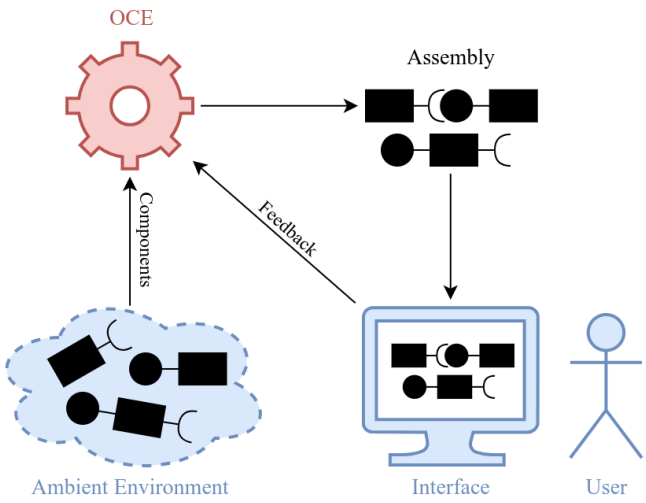


Fig. 1: Iterative assembly, feedback, and learning cycle of OCE

than enforcing predefined assemblies that may become quickly outdated due to changing preferences or dynamic availability of components. The idea is to take what is available and construct the best possible service in a given context.

The operational principle of OCE follows the iterative cycle illustrated in Figure 1. To construct an assembly, OCE gathers information about components present in the environment. The proposed assembly is then presented to the user, who can accept, modify, or reject it. If not rejected, this interaction triggers both deployment in the environment and the learning process of OCE.

Internally, OCE relies on a multi-agent reinforcement learning mechanism [21], where agents represent available services and collaborate to propose optimal assemblies based on their knowledge. Agents aim to maximize positive reward computed from the user feedback (acceptance of assemblies by the user) while balancing the exploitation of known effective combinations with the exploration of new ones. The key feature of OCE is its ability to adapt dynamically to changing environments and user preferences, integrating humans in the learning loop without requiring explicit specification of all preferences [15].

B. Simulated Ambient Environment

To interact with OCE and support our user studies, we developed a 2D simulated ambient environment. This simulation abstracts away the challenges of deploying real-world devices while allowing live interaction with OCE in a Web-based interface accessible from any browser. It provides a simplified testbed to focus on the interaction logic and learning behavior of the system. The supplementary material [22] includes a video demonstrating interaction with OCE in this environment.

The environment resembles a 2D game in which the user controls an avatar and can interact with various components, such as doors, lights, sensors, and smartphone-based apps. These components are managed by OCE enabling users to request assembly propositions, modify them, and accept or reject them, which triggers the corresponding learning and deployment processes. The environment also includes the possibility to create simple goals, such as opening a door or displaying a temperature, to create meaningful tasks for our user studies.

Figure 2 illustrates a typical interaction in the simulated environment. Users can explore a space populated with diverse components (Figure 2a). When they need to interact with the components, they can request OCE to propose an assembly, which is displayed as components, nodes, and service links (Figure 2b). The users can modify connections by dragging them to adjust the assembly to their goals then accept, or reject the assembly (Figure 2c). Accepted assemblies are deployed in the environment and provide feedback to OCE's learning agents, while rejected ones prompt further exploration. Once deployed, users can directly interact with components, e.g., clicking a smartphone button to open a door (Figure 2d), completing the perception-action-learning cycle.

C. Granular Feedback Integration in OCE

Based on design recommendations from [13] on human–AI interaction in AmI systems, we extended OCE to allow users to provide feedback at multiple levels of granularity. This complements the baseline OCE described earlier by enabling a more nuanced understanding of user preferences, which can lead to better recommendations and improved system performance.

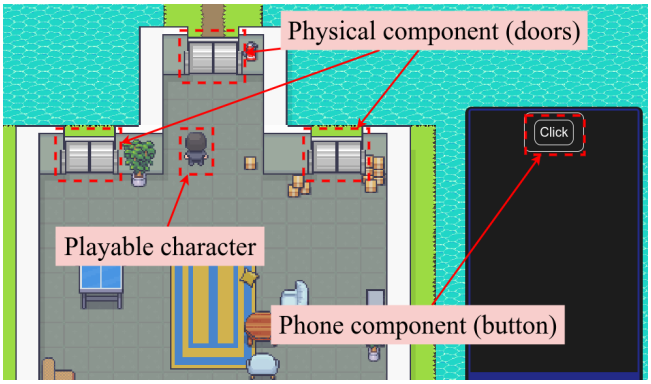
Users can provide positive or negative feedback on four levels: individual services, connections between components, entire components, or complete assemblies (Figure 3). Each feedback signal is interpreted by OCE's multi-agent system to adjust the likelihood that the corresponding service, connection, component, or assembly appears in future proposals. This flexible feedback mechanism allows the system to finely adapt to changing user needs and the dynamic ambient environment.

In the simulated environment, the feedback is represented by thumbs-up and thumbs-down icons that appear in the tooltips for each level of granularity upon mouse hover (Figure 3). Clicking these icons updates the MAS knowledge, reinforcing or diminishing the associated elements in future assembly propositions.

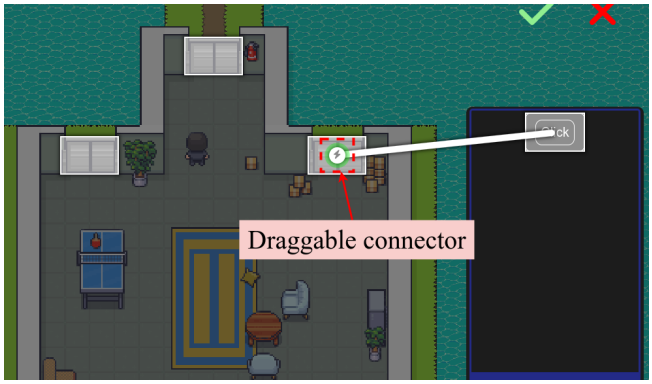
In the studies presented in this paper, we compare this granular-feedback version of OCE (version B) with the baseline version without granular feedback (version A).

V. Study Designs

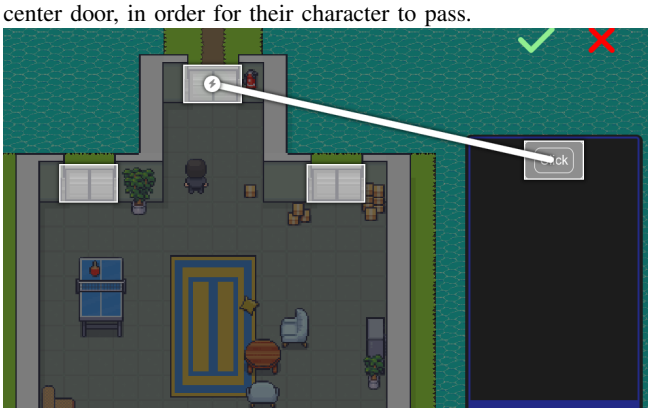
We conducted two between-subjects user studies to investigate the impact of granular feedback in an adaptive AmI system. The first study focused on evaluating OCE's usability (Study 1), while the second examined the ML accuracy of OCE's recommendation mechanism (Study 2). In both studies, we compared the baseline version of OCE (condition A) with the version incorporating granular feedback mechanisms (condition B), as described in Section IV.



(a) The user must connect their phone's button to the environment's center door, in order for their character to pass.



(b) OCE proposes to connect the button with the door on the right of the environment.



(c) The user corrects the engine by dragging the connector to the correct component.

(d) The user clicks on the button to open the door. They can advance to the next step.

Fig. 2: Simple interaction with the simulated ambient environment. The user controls an avatar and can interact with components, and with OCE in order to act out AmI use cases.

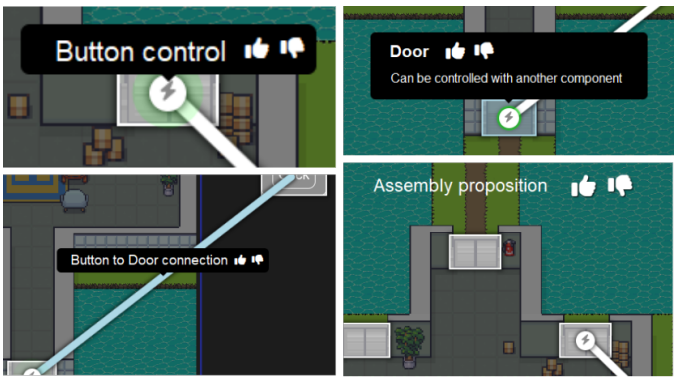


Fig. 3: The four possible levels of granularity for user feedback, along with the associated screenshots. A user can provide feedback on a service, a connection, a component, or an assembly by interacting with the corresponding "thumbs up/down" buttons.

The studies were conducted over a period of five months. In both studies, participants completed a demographic questionnaire followed by a tutorial designed to familiarize them with OCE and its simulated AmI environment. They then performed a series of tasks and goals with the assistance of OCE, before completing a post-study questionnaire assessing relevant aspects of usability (Study 1) or system behavior (Study 2). Participants were also given the opportunity to provide their opinion through an open-ended comment section. In parallel, interaction logs and system data were collected on the server side to support our analysis. Each participant interacted with an independent instance of OCE's multi-agent system, which was initialized without any prior information about the participant or the ambient environment.

Rationale for Two Studies: We deliberately separated the evaluation into two studies to address distinct methodological requirements associated with usability assessment and learning performance evaluation. Measuring usability in adaptive systems typically requires providing users with a high degree of freedom and ecological validity, allowing them to interact naturally with the system [23]. However, this flexibility makes it difficult to rigorously assess learning performance, as user goals and preferences may vary widely and cannot be explicitly specified.

Conversely, evaluating learning performance necessitates a more constrained experimental setting in which task objectives and expected system behaviors are well defined, enabling

objective measurements. Such constraints, however, may limit the realism of the interaction and reduce the validity of usability measurements. To address this tension, Study 1 adopts a more open-ended, realistic interaction scenario suitable for assessing usability, while Study 2 employs a more structured task design that enables precise evaluation of learning behavior and system performance.

Each study involved different task sets, goals, and evaluation metrics aligned with their respective research questions. We provide, as supplementary material [22], the complete list of tasks and goals that participants were required to do in each study.

A. Study 1 – Usability

The goal of Study 1 was to simulate a realistic interaction scenario in a smart environment, in this case a smart campus. Participants were asked to complete a series of tasks typical for campus users, such as conducting classes, navigating between buildings, and using campus facilities. To preserve ecological validity, participants were free to choose the order in which they completed tasks, allowing them to interact with the environment naturally rather than following a rigid sequence.

B. Study 2 – Machine Learning Performance

Complementing Study 1, Study 2 was conducted in a more constrained environment to allow precise control over the "correct" component assemblies at each step of the experiment. The setting simulates a smart home, focusing on typical morning and evening routines, such as preparing meals, setting alarms, and managing household devices.

Participants completed five series of morning and evening routines (referred to as "experiment days" in the following) to provide sufficient repetitions for OCE's machine learning mechanism to adapt to their preferences. To maintain ecological validity, no two experiment days were identical; for instance, some components malfunctioned or user preferences shifted between days, simulating the dynamic nature of real-world environments.

C. Participants

The experiments were conducted online using a web browser with no assistance provided to participants other than the tutorial. The URL to the experimentation platform was made publicly available and participants were recruited through social media and various academic mailing lists Participation criteria included being a French speaker, as the experiments were conducted exclusively in French, and having access to a computer with a keyboard and mouse or trackpad, since the simulated environment was not compatible with smartphones. No compensation (monetary or otherwise) was provided for participation.

Study 1 involved $N_1 = 61$ participants and was conducted over a period of four months. For Study 2, as we targeted experienced users in order to focus on system performance, the experiments were conducted primarily with participants who had already completed Study 1. As a consequence of time constraints, Study 2 lasted only one month, resulting in $N_2 = 12$ participants. In both studies, participants were randomly assigned to condition A or B, ensuring roughly equal numbers in each condition (for S1: 31 in A and 30 in B; for S2: 6 in A and 6 in B).

In both studies, most participants were aged between 18 and 35 years (S1: 78.7% – S2: 58.3%), held a Master's degree or higher (S1: 55.7% – S2: 75.0%), and were from the computer science field (S1: 77.0% – S2: 91.7%). Additionally, the vast majority reported playing video games at least occasionally (S1: 96.7% – S2: 91.6%), reflecting the need for a population capable of navigating a 2D environment with minimal assistance. The supplementary material [22] includes a detailed breakdown of the demographic questionnaires used in each user study.

D. Measures

Here, we describe the measures collected across our studies, organized according to our research questions: usability measures for RQ1, machine learning performance measures for RQ2, and combined measures for assessing trade-offs for RQ3.

a) Usability Measure: To address **RQ1**, usability was assessed in Study 1 using the *USE Questionnaire* [24], which was translated into French for this study. The USE questionnaire consists of 30 items rated on a 1–7 Likert scale and evaluates four dimensions of user-perceived interaction: *usefulness*, *ease of use*, *ease of learning*, and *satisfaction*. Participants had the option to select "Not Applicable" (NA) for any item, and responses marked as NA were excluded from subsequent statistical analyses. Additionally, three items (items 7, 9, and 26) were formulated in a negative form to mitigate potential response bias. The supplementary material [22] contains the full questionnaires in both their original and translated versions.

Several standardized instruments are available for evaluating user experience. For instance, the *System Usability Scale* (SUS) [25] is widely used and concise (10 items). Comparative analyses of usability questionnaires [26] indicate that the USE questionnaire is among the few instruments that systematically measure both usability and user satisfaction, which is an additional insight into general system performance [27]. For these reasons, the USE questionnaire was selected as the primary instrument for assessing participants' perceived usability in this study.

b) Machine Learning Performance Measures: The controlled nature of Study 2 allowed us to define a unique correct output (OCE assembly) for each step, which served as the ground truth. We then compared OCE's recommendations against this ground truth for each recommendation and computed an accuracy score:

$$\text{Accuracy} = \frac{\text{Count of Correct Recommendations}}{\text{Count of Correct + Incorrect Recommendations}}$$

To capture the evolution of OCE's learning over time, we computed two measures for each participant: (1) the average

accuracy across all experiment days, and (2) the maximum accuracy achieved on any single experiment day. These metrics allow us to assess both overall learning performance and the system's best-case adaptation capability.

c) Combined Measures: To address **RQ3**, we combine the results from Study 1 (usability) and Study 2 (ML performance) to evaluate potential trade-offs between the two objectives. For each OCE version (A and B), we compute an average usability score from the USE questionnaires and an average accuracy score from the machine learning performance measures.

Usability and ML performance can be viewed as competing objectives, and the trade-off can be conceptualized similarly to a multi-objective optimization problem [28]. Specifically, we define a local Pareto slope representing the accuracy gain per unit of usability change between two configurations:

$$s = \frac{|accuracy_B - accuracy_A|}{|usability_B - usability_A|}$$

This measure allows us to directly compare interface alternatives in terms of their relative benefit for learning performance versus potential cost in usability.

Because usability is measured on a bounded Likert scale and the relationship between usability and performance may be nonlinear, this slope should not be interpreted as a universal or absolute metric. Instead, it serves as a descriptive, local indicator for contrasting design options under identical experimental conditions.

E. Ethical Considerations

Although formal Institutional Review Board approval was not required for this study, we adhered to established ethical guidelines for human-subject research. Participants provided informed consent prior to participation and were informed of their right to withdraw at any time. No sensitive or personally identifiable data were collected, and all responses were anonymized before analysis.

VI. Results

Across the two studies, each participant completed an average of approximately 70 OCE cycles, where each cycle corresponds to OCE making an assembly recommendation that the participant either accepted, modified, or rejected. On average, participants spent around 40 minutes completing the experiment. In the granular feedback condition (Group B), participants used the granular feedback feature an average of 11 times.

The results shown in Figures 4, 5, and 6 are presented as boxplots, with the number of participants in each group indicated and the mean scores represented by dashed lines. Statistical significance for these results was assessed using Student's t-test, Welch's t-test, or the Mann–Whitney U test, as appropriate. Each test yields a p-value, denoted as p in the respective figures. A p-value below 0.05 (respectively 0.01) is considered statistically significant (respectively highly significant) and is indicated by a single asterisk * (respectively double asterisks **).

The supplementary material [22] contains the raw, anonymized data collected from the experimental platform, along with all scripts required to generate the figures and perform the statistical analyses (e.g., t-tests) reported in this section.

A. RQ1 – Usability

Before interpreting the USE questionnaire results, we first assessed its validity and reliability. Following the approach advocated by Gao et al. [29], participants with more than 5 NA responses in the questionnaire were excluded from the data analysis. This resulted in 4 exclusions in group A, resulting in 27 questionnaires taken into account for this study.

Regarding validity (i.e., whether the questionnaire measures what it is intended to measure), Gao et al. [29] conducted a psychometric evaluation of the USE questionnaire and concluded that it is a valid instrument, based on comparisons with the SUS [25] across various applications.

For reliability (i.e., the extent to which the instrument produces stable and consistent measurements over time [30]), we computed Cronbach's alpha for each USE questionnaire dimension to assess internal consistency in the translated French version. The resulting alphas were: Usefulness $\alpha = 0.92$, Ease of Use $\alpha = 0.89$, Ease of Learning $\alpha = 0.87$, and Satisfaction $\alpha = 0.94$, indicating good reliability across all dimensions.

a) Per-dimension results: Figure 4 shows the distribution of participant responses for each USE questionnaire dimension. Across all dimensions, there was a decrease in scores from version A to version B. The p-values indicate that this decrease was statistically significant for the *Satisfaction* and *Ease of Use* dimensions, whereas the reductions in *Ease of Learning* and *Usefulness* were smaller and not statistically significant.

These results indicate that introducing the granular feedback option had a measurable impact on participants' perceived satisfaction and ease of use of OCE. Several participants in group B commented on the intrusive nature of the feedback prompts, for example: "Too many pop-ups; it can disrupt the user experience." Nevertheless, participants still generally found OCE useful and easy to learn, with one participant noting, "I really like this interaction with the user."

b) Overall usability: Figure 5 presents a single usability score per participant, calculated by averaging the scores of the 30 USE items. The overall results confirm the per-dimension findings: usability decreased significantly from version A to version B. Participants attributed this primarily to the graphical interface for providing granular feedback, which they found to be overly cluttered; as one participant commented, "The cause [of my frustration] is that the graphical interface is too easily overloaded."

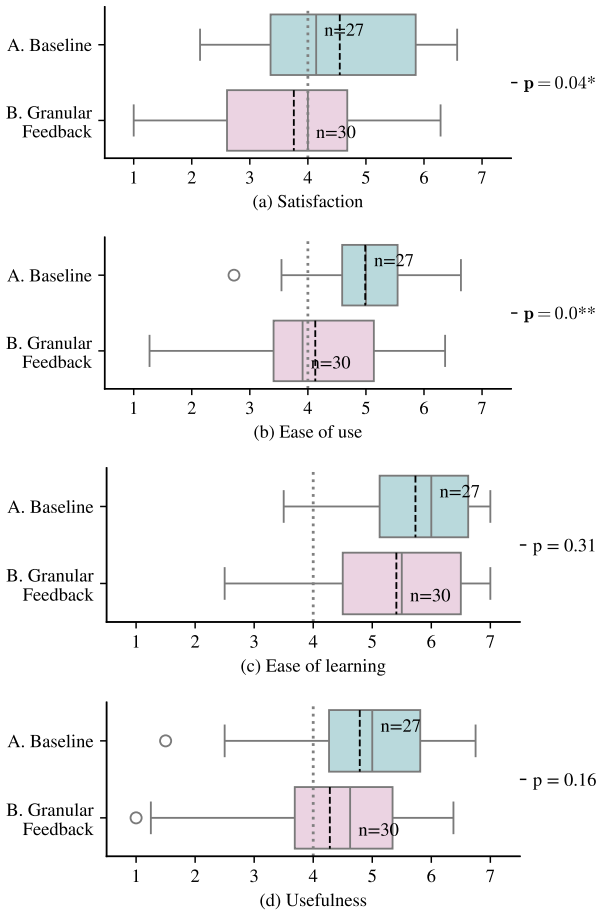


Fig. 4: RQ1. Participants' answers to the USE questionnaire for each dimension.

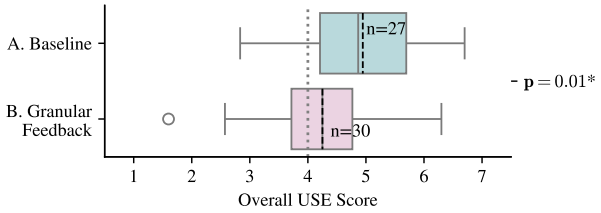


Fig. 5: RQ1. Usability of OCE across the tested versions.

RQ1 - How does granular feedback affect OCE's perceived usability?

The introduction of granular feedback led to a measurable decrease in participants' perceived usability, with significant reductions in satisfaction and ease of use, while usefulness and ease of learning were largely unaffected.

B. RQ2 – Learning Performance

Figure 6 reports the average accuracy of OCE during Study 2 for each participant group, as well as the maximum accuracy achieved on any single experiment day (approximately 20 OCE cycles).

These results allow us to observe the evolution of OCE's learning capabilities over time. Average accuracies were initially low (below 0.5), reflecting the fact that OCE started with no prior knowledge of participants' preferences. In addition, the tasks and component assemblies were deliberately designed to be complex, providing a realistic challenge for OCE's multi-agent system. After several repetitions, however, accuracy improved substantially as shown in the maximum reported, with one participant reaching a maximum accuracy of 1.0. While this is an exception, overall this demonstrates that OCE successfully adapted to participant preferences over the course of the experiment.

Comparing the two groups, participants in the granular feedback condition (B) achieved significantly higher accuracy than those in the baseline condition (A). This indicates that providing granular feedback substantially enhanced OCE's learning performance: participants were able to communicate their preferences more effectively, which allowed the system to generate more accurate assembly recommendations and complete tasks more successfully.

RQ2 - How does granular feedback affect OCE's learning performance?

Providing granular feedback significantly improved OCE's learning performance, allowing the system to adapt more quickly and accurately to participants' preferences compared to the baseline condition.

C. RQ3 – Trade-off

Figure 7 illustrates the trade-off between usability and maximum accuracy achieved by OCE for the Baseline (A) and Granular Feedback (B) conditions. For each condition, we computed a mean usability score and a mean accuracy score by averaging the corresponding measures across participants expressed as percentages. Ellipses summarize the variability of the accuracy and usability values and are included for visualization purposes only. The local Pareto slope representing the trade-off is visualized with an arrow connecting the two conditions.

This visualization shows that introducing granular feedback leads to a gain in system performance at the cost of decreased usability. The slope of the trade-off line, calculated as $s = \Delta_{\text{accuracy}}/\Delta_{\text{usability}} \simeq 0.78$, indicates that each one-percent decrease in usability corresponds to a 0.78 percent increase in accuracy, meaning that the increase in performance is rather expensive in term of usability.

While this slope is specific to our experimental setting and should not be interpreted as a universal law (particularly when these measures originate from different studies) it illustrates a measurable trade-off between learning performance and usability. In our study, the trade-off is tilted toward usability: the performance gains afforded by granular feedback come at a substantial cost to the user experience.

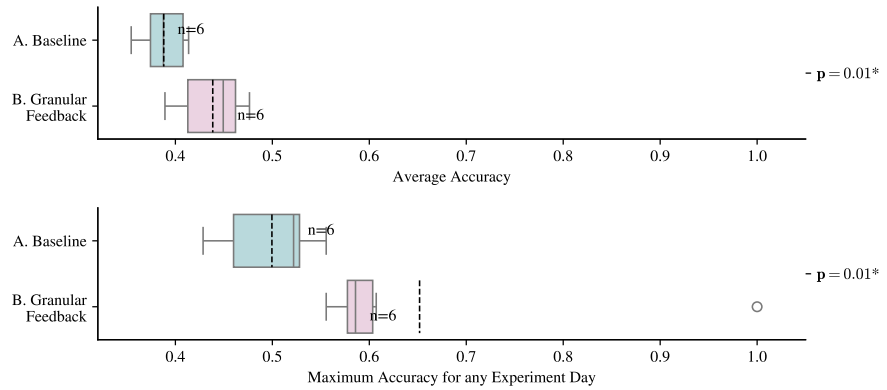


Fig. 6: RQ2. Average and Maximum Accuracy of OCE across the different versions.

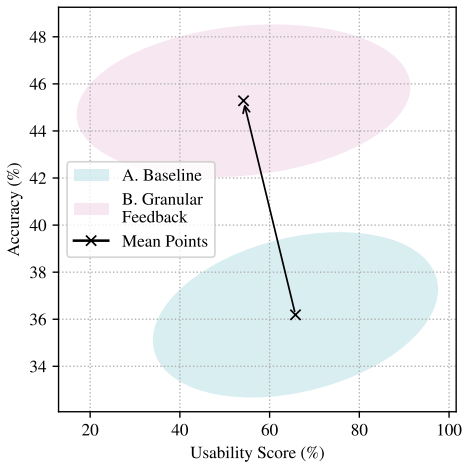


Fig. 7: RQ3. Usability–Performance Trade-off across Study 1 and Study 2. Ellipses indicate the overall spread of the data along each axis

RQ3 - What trade-offs arise between usability and learning with granular feedback?

The introduction of granular feedback creates a clear trade-off between learning performance and usability: it improves OCE's accuracy but at a measurable cost to user experience, with the trade-off in our study being more heavily weighted toward usability loss.

VII. Discussion

The results of our studies provide empirical insights into how granular feedback affects both user experience and machine learning in Human–AmI systems. They highlight a clear trade-off between usability and performance and suggest design considerations for adaptive interfaces that rely on human-in-the-loop feedback.

A. Limitations and Threats to Validity

We distinguish between internal validity, factors that may have influenced our measurements, and external validity, factors that may limit the generalizability of our results.

a) Internal Validity: We modified the USE questionnaire through translation into French and by rewording three items in a negative form. These changes may have introduced biases affecting construct validity. However, internal consistency was assessed using Cronbach's alpha, which showed good reliability across all dimensions, suggesting that the translated questionnaire remained a coherent measurement instrument.

Regarding Study 2, although task execution followed standardized routines in the form of experiment days (see Section V-B), the variations between those days (i.e., which component was set up to fail or reappear at what time) did influence OCE's learning dynamics and performance. These variations affected the ML accuracy measurements negatively, but were intentionally introduced to reflect realistic challenges in AmI systems, thereby increasing ecological validity.

Finally, prior exposure bias may have influenced the results. Given the demographic composition of our participants, some had prior experience with similar tasks, computer science, or video games. This affected how quickly participants learned the interface or how effectively they used the granular feedback mechanism, impacting ML performance outcomes.

Overall, while these factors may have influenced absolute measured values, they do not undermine the core comparative findings of the study, which focus on differences between interaction conditions under otherwise comparable settings.

b) External Validity: Several factors limit the generalizability of our findings. Our participant pool was relatively homogeneous, consisting predominantly of young, French-speaking individuals with computer science backgrounds. Interactions took place in simulated environments, over relatively short experimental sessions (around 40 minutes per participant) and with a limited set of tasks.

Moreover, the experiments were conducted in simulated environments, which necessarily simplify real-world Ambient Intelligence systems. In real deployments, additional sources of complexity–such as environmental distractions, long-term usage, and evolving user goals–may affect both perceived usability and system learning outcomes.

Finally, the results are tied to the specific interface, feedback modality, and timing implemented in OCE. To disentangle the general impact of granular feedback mechanisms from their particular implementation, future work should explore multiple interface designs and interaction strategies. While alternative implementations may mitigate some of the observed usability costs, our results nonetheless illustrate a concrete instance of the usability–performance trade-offs inherent to granular feedback in Human–AI and Human–AmI interaction.

B. Implications and Future Works

While the tension between usability and system performance is well documented in Human–AI research [14], [17], [18], our results show that this trade-off also persists in adaptive Ambient Intelligence systems and can emerge even from relatively modest interaction design changes.

a) For Opportunistic Software Composition: Our study of OCE's interaction design and its impact on usability and system performance provides insights into the design of opportunistic software composition systems more broadly. While this result illustrates a clear usability–learning performance trade-off, it also suggests that not all levels of granularity are equivalent. Soliciting feedback only at specific levels of abstraction (e.g., at the component level rather all four levels) may offer a more favorable balance between usability and learning performance, effectively pushing the Pareto frontier.

Moreover, this work opens avenues toward deploying opportunistic software composition systems in real-world Ambient Intelligence environments. Studying such systems in real-world settings—where user preferences, available components, and environmental constraints evolve over time—constitutes an important direction for future research.

b) For Adaptive Human-AI Systems: Our findings indicate that granular feedback can substantially improve system learning while simultaneously degrading perceived usability. This highlights that user feedback should be treated as a strategic resource rather than a constant interaction requirement. Not all steps in a task necessarily warrant detailed feedback, and designers should carefully consider when to request granular input, at which moments in the interaction, and at what level of detail. In the context of ambient intelligence, constant interaction can undermine the goal of unobtrusive support by increasing cognitive load and interrupting user activities. Instead, feedback mechanisms should be designed to maximize informational value while minimizing disruption—for example, by triggering feedback requests at moments of uncertainty, system failure, or model retraining phases. Poorly timed or overly frequent feedback requests may impose cognitive and interaction costs that outweigh their performance benefits, as hinted by our answer to RQ3.

These observations are consistent with prior work on feedback granularity. For example, Peska et al. [31] show that overly coarse feedback may be insufficiently informative for learning systems, while excessively fine-grained feedback is often overlooked or underutilized by users. Together with our findings, this suggests that effective feedback design requires balancing informativeness for the system with effort and attention demands on the user.

An alternative direction explored in the literature is the collection of implicit feedback, such as signals derived from mouse movements or interaction traces [32]. Peska et al. [31] frame granular feedback as a more explicit and often complementary alternative to such implicit signals. While implicit feedback can reduce user burden, it is noisy and difficult to interpret, potentially leading to erroneous inferences [33]. More broadly, integrating human feedback into machine learning systems remains challenging: as noted by Nguyen et al. [34], human feedback is subjective, varies across users, and is not always directly translatable into machine-interpretable signals.

c) For AmI Study Design: Beyond design implications, our work offers methodological insights for future research on Human–AmI systems. First, the two-study setup proved effective despite its added complexity. Measuring usability in a realistic, open-ended scenario while assessing learning performance under more controlled conditions allowed us to capture complementary aspects of the interaction. Second, starting with realistic yet constrained simulations provided a practical testbed for studying adaptive behavior. This allowed us to focus on the design of feedback and interaction itself, without addressing other practical challenges covered elsewhere in the literature, such as sensor data acquisition [35], or data safety [36] in AmI systems.

This combination of experimental design choices can serve as a useful starting point for future work in Human–AI and Ambient Intelligence research. Future studies should extend this approach by exploring alternative interface designs, feedback modalities, and longer-term deployments, in order to better characterize and potentially mitigate the usability–performance trade-offs identified in this work.

VIII. Conclusion

This paper investigated how introducing granular user feedback in an adaptive Ambient Intelligence system affects both user-perceived usability and machine learning performance. Through two complementary user studies, we evaluated a baseline version of an opportunistic software composition prototype, OCE, against a version enabling explicit, fine-grained feedback. The first study focused on usability, while the second examined learning performance under more controlled conditions. Together, our results show that granular feedback significantly improves system learning performance, enabling OCE to better adapt to user preferences over time. At the same time, this improvement comes at a measurable cost to usability, particularly in terms of user satisfaction and ease of use. These findings highlight that richer feedback mechanisms are not universally beneficial and must be carefully calibrated within the interaction flow.

Beyond the specific case of OCE, this work contributes to a growing body of research on agentic and adaptive AI systems that rely on human input to guide learning and decision-making. As AI agents become increasingly proactive, persistent, and embedded in everyday environments, the question is

no longer whether humans should provide feedback, but how, when, and at what level of granularity such feedback should be solicited. This work examines the trade-offs between usability and ML performance in this context, a topic that is largely unexplored in the literature. Our results suggest that feedback should be treated as a limited and valuable resource, rather than a default interaction pattern.

For the broader agentic AI community, this work underscores the importance of jointly evaluating learning performance and human experience when designing adaptive systems. Optimizing agents solely for learning efficiency risks degrading user experience in ways that may ultimately undermine long-term adoption and trust. We argue that future research should further explore adaptive feedback strategies, and alternative feedback modalities to better understand how agentic systems can learn effectively while remaining usable by their users. In this sense, our work represents a step toward more human-centered approaches to designing and evaluating learning agents in Ambient Intelligence and beyond.

REFERENCES

[1] C. Diebel, M. Goutier, M. Adam, and A. Benlian, "When AI-Based Agents Are Proactive: Implications for Competence and System Satisfaction in Human–AI Collaboration," *Business & Information Systems Engineering*, pp. 1–20, 2025.

[2] K. Yuan, Y. Huang, L. Guo, H. Chen, and J. Chen, "Human feedback enhanced autonomous intelligent systems: a perspective from intelligent driving," *Autonomous Intelligent Systems*, vol. 4, no. 1, p. 9, 2024.

[3] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," 2017.

[4] A. Gelb, N. Mittal, and A. Mukherjee, "Towards real-time governance: Using digital feedback to improve," *Center for Global Development. https://www. cgdev. org/publication/towards-real-time-governance-using-digital-feedback-improve-service-voice*, 2019.

[5] H. Peleg, S. Shoham, and E. Yahav, "Programming not only by example," in *Proceedings of the 40th International Conference on Software Engineering*, 2018, pp. 1114–1124.

[6] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, and K. Inkpen, "Guidelines for human-AI interaction," in *Proc. of the 2019 CHI conf. on human factors in computing systems*, 2019, pp. 1–13.

[7] F. Sadri, "Ambient intelligence: A survey," *ACM Computing Surveys*, vol. 43, no. 4, pp. 1–66, 2011.

[8] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos, "A survey on ambient intelligence in healthcare," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2470–2494, 2013.

[9] S. Makonin, L. Bartram, and F. Popowich, "A smarter smart home: Case studies of ambient intelligence," *IEEE pervasive computing*, vol. 12, no. 1, pp. 58–66, 2012.

[10] S. D. Nagowah, H. Ben Sta, and B. Gobin-Rahimbux, "A systematic literature review on semantic models for IoT-enabled smart campus," *Applied ontology*, vol. 16, no. 1, pp. 27–53, 2021.

[11] A. B. Karami, A. Fleury, J. Boonaert, and S. Lecoeuche, "User in the loop: Adaptive smart homes exploiting user feedback—state of the art and future directions," *Information*, vol. 7, no. 2, p. 35, 2016.

[12] Y. Park and J. Han, "Smart home advancements for health care and beyond: Systematic review of two decades of user-centric innovation," *Journal of Medical Internet Research*, vol. 27, p. e62793, 2025.

[13] K. Delcourt, S. Trouilhet, J.-P. Arcangeli, and F. Adreit, "The human in interactive machine learning: analysis and perspectives for ambient intelligence," *Journal of Artificial Intelligence Research*, vol. 81, pp. 263–305, 2024.

[14] M. Fan, X. Yang, T. Yu, Q. V. Liao, and J. Zhao, "Human-ai collaboration for ux evaluation: effects of explanation and synchronization," *Proceedings of the ACM on human-computer interaction*, vol. 6, no. CSCW1, pp. 1–32, 2022.

[15] W. Younes, S. Trouilhet, F. Adreit, and J.-P. Arcangeli, "Agent-mediated application emergence through reinforcement learning from user feedback," in *29th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises*, 2020, pp. 3–8.

[16] H. Yu, R. M. Aronson, K. H. Allen, and E. S. Short, "From "thumbs up" to "10 out of 10": Reconsidering scalar feedback in interactive reinforcement learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4121–4128.

[17] D. Honeycutt, M. Nourani, and E. Ragan, "Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy," in *Proceedings of the AAAI Conf. on Human Computation and Crowdsourcing*, vol. 8, 2020, pp. 63–72.

[18] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability," *International Journal of Information Management*, vol. 69, p. 102538, 2023.

[19] J. Alves Lino, B. Salem, and M. Rauterberg, "Responsive environments: User experiences for ambient intelligence," *Journal of ambient intelligence and smart environments*, vol. 2, no. 4, pp. 347–367, 2010.

[20] K. Delcourt, F. Adreit, J.-P. Arcangeli, K. Hacid, S. Trouilhet, and W. Younes, "Automatic and Intelligent Composition of Pervasive Applications - Demonstration," in *19th IEEE Int. Conf. on Pervasive Computing and Communications (PerCom 2021)*, Kassel (virtual), Germany, Mar. 2021.

[21] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[22] K. Delcourt, M. Renard, S. Trouilhet, and J.-P. Arcangeli, "Granular feedback in human–ai interaction: Exploring trade-offs in ambient intelligence (supplementary materials)," Jan. 2026. [Online]. Available: https://doi.org/10.5281/zenodo.18396807

[23] L. Van Velsen, T. Van Der Geest, R. Klaassen, and M. Steehouder, "User-centered evaluation of adaptive and adaptable systems: a literature review," *The knowledge engineering review*, vol. 23, no. 3, pp. 261–281, 2008.

[24] A. M. Lund, "Measuring usability with the use questionnaire," *Usability interface*, vol. 8, no. 2, pp. 3–6, 2001.

[25] J. Brooke, "Sus: a retrospective," *Journal of usability studies*, vol. 8, no. 2, pp. 29–40, 2013.

[26] A. Assila, H. Ezzedine *et al.*, "Standardized usability questionnaires: Features and quality focus," *Electronic Journal of Computer Science and Information Technology*, vol. 6, no. 1, 2016.

[27] J. R. Griffiths, F. Johnson, and R. J. Hartley, "User satisfaction as a measure of system performance," *Journal of Librarianship and Information Science*, vol. 39, no. 3, pp. 142–152, 2007.

[28] K. Deb, "Multi-objective optimisation using evolutionary algorithms: an introduction," in *Multi-objective evolutionary optimisation for product design and manufacturing*. Springer, 2011, pp. 3–34.

[29] M. Gao, P. Kortum, and F. Oswald, "Psychometric evaluation of the use (usefulness, satisfaction, and ease of use) questionnaire for reliability and validity," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 62, no. 1. Sage Publications, Los Angeles, CA, 2018, pp. 1414–1418.

[30] E. G. Carmines and R. A. Zeller, *Reliability and validity assessment*. Thousand Oaks, Ca., USA: Sage Publications, 1979.

[31] L. Peska and S. Balcar, "The effect of feedback granularity on recommender systems performance," in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 586–591.

[32] Y. Liu, Y. Chen, J. Tang, J. Sun, M. Zhang, S. Ma, and X. Zhu, "Different users, different opinions: Predicting search satisfaction with mouse movement information," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 493–502.

[33] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search," *ACM Transactions on Information Systems (TOIS)*, vol. 25, no. 2, pp. 7–es, 2007.

[34] K. Nguyen, H. Daumé III, and J. Boyd-Graber, "Reinforcement learning for bandit neural machine translation with simulated human feedback," *arXiv preprint arXiv:1707.07402*, 2017.

[35] C. Wu, D. Peng, Y. Lu, Y. Zou, Z. Liu, H. Onoda, H. Washizaki, W. Kameyama, and J. Liu, "Cfdiff: A diffusion-based generative framework for efficient multi-physical field prediction in smart iot," *IEEE Internet of Things Journal*, 2025.

[36] H. Bian, W. Zhang, and C. K. Chang, “Situ-oracle: A learning-based situation analysis servicing framework for biot systems,” in *2023 IEEE International Conference on Software Services Engineering (SSE)*. IEEE, 2023, pp. 159–169.