EmoReflex: An AI-powered Emotion-centric Developer Insights Platform

Kashumi Madampe[1*], John Grundy[1], Minh Nguyen[1], Ellen Welstead-Cloud[1], Vinh Tuan Huynh[1], Linh Doan[1], Wlliam Lay[1], Sayed Hashim[1]

[1*]Department of Software Systems and Cybersecurity, Monash University, Wellington Road, Clayton, 3800, Victoria, Australia.

*Corresponding author(s). E-mail(s): kashumi.madampe@monash.edu;
Contributing authors: john.grundy@monash.edu;
tngu0235@student.monash.edu; ewel0002@student.monash.edu;
vhuy0004@student.monash.edu; ndoa0012@student.monash.edu;
whua0022@student.monash.edu; shas0010@student.monash.edu;

Abstract

There has been great interest in better understanding software engineer emotions during development. But how to do this? We built a prototype AI-powered Emotion-centric Developer Insights Platform, EmoReflex, to support developers to report and reflect how they feel when working on various tasks across different metrics. It also assists their managers to get insights into their team's emotional health, and provides them with recommendations to guide them handle the team's emotional wellbeing. We present our tool prototype and evaluation results generated by a user study conducted with two user groups consisting of twenty developers and twenty managers. We present some design implications derived from our user study that can be used to inform design decisions in emotion-centric software development tools.

Keywords: emotions, emotional intelligence, affects, requirements, changes, human factors, software engineering, software teams, socio-technical grounded theory, agile, well-being, workplace awareness, productivity, team goals, developer tools

1 Introduction

The management of software teams is not only limited to technical task management, but also critically needs to encompass skilled people management. For a team to better perform and be productive, it is necessary for developers themselves to have awareness about their performance, and also for their managers to have awareness of the team well-being [1]. Increasingly it has been recognised that *emotions* play a crucial role when working in a team at both team and individual levels. Emotions act as behavioural motivators [2], and have direct linkages to cognition [2], productivity [2–4], and decision-making [5].

Software teams use a variety of tools to manage their projects. These tools tend to focus on technical task management, and focus on better managing the projects but not necessarily always on better managing the people. We believe that project/task management tools need to give attention to the human aspects of the developers. However, one area of still under-supported care of software developers is their *emotional well-being* [6]. In previous work we built Emotimonitor [6], a Trello plugin for self-reporting and managing of emotions of agile teams. Emotiomonitor had the functions of rating against a set of specific emotions for the Trello cards for the developers, and a dashboard for the managers to see the emotions for the Trello cards in any given time period. Following up on this work, we designed and built EmoReflex, a novel 'developer insights' platform that is emotion-centric, to promote incorporating emotional well-being/intelligence in working in software teams and managing software teams. EmoReflex is a much improved and extended version of Emotimonitor. It allows the developers not only to rate against the emotions, but also to rate against metrics, and add reflections (text), and the managers to configure the Trello plugin for emotions and metrics, view emotion and metrics summary by task, overall, any time period, and also to receive recommendations generated by AI. In this paper, we present the key motivation and features of EmoReflex and early feedback of the tool from 40 potential users. We first present a usage scenario of EmoReflex in section 2, then the system design and implementation of it in section 3. Then we present the feedback we gathered from 40 developers and managers collectively in section 4. Finally we discuss the implications of EmoReflex in section 5.

2 Motivating Scenario

Imagine a software development team comprising several developers and a manager. The team uses Trello as their task management tool. The developers are assigned with various tasks across the project. The developers find tasks assigned to them sometimes challenging, sometimes not, sometimes they feel negative emotions, sometimes they feel positive emotions when working on them. They also want to reflect how they feel when working on the tasks. Some developers find it difficult to communicate directly to their manager about how they feel. What if they have a tool which supports them to accomplish the mentioned?

On the other hand, the team manager wants to obtain a summary of the key tasks the team is performing and wants to get a snapshot of the tasks across various metrics they would like to define. For example, the manager wants to know how complex a task

is for the team, how difficult could the tasks be. The manager also wants to know how the team feels when working on the tasks assigned to them. The manager believes that being compassionate about the team is necessary for better managing them. However, they feel challenging to access some of the team members to discuss how they feel. What if they have a tool which supports them to achieve the mentioned?

3 Our EmoReflex Approach

To address the scenario above we built a prototype we call EmoReflex. EmoReflex is intended to support agile software teams using Trello boards to manage the tasks of their development projects including emotional impacts of tasks. We define a software team as a group consisting of a manager and a set of developers – the two user groups of EmoReflex. In practical terms, the members of the team could also include business analysts, product designers, user researchers, etc., and they can be included in the user group we define as "developers" in this work.

3.1 User Interfaces

The Trello plugin is available for the developers (Fig. 1), and the dashboard is only available for the manager (Fig. 2). i.e., from the user interface point of view, the developers do not have the dashboard view, but only the manager.

The developer view is for the developers who have activated the plugin at their end. If the manager plays a developer role within the team, they can have their own emotions and metrics tracked too using the Trello plugin activated on their Trello end. This is further elaborated in the following subsections.

3.1.1 Trello Plugin

A manager configures at the dashboard end a set of emojis and metrics that appear on the Trello plugin for the developers to use. Fig. 1 shows a single metric and the user interactions. If multiple metrics are configured, multiple metric boxes appear on the screen of the Trello plugin.

3.1.2 Dashboard

The dashboard allows the manager to configure EmoReflex, view team summary, and receive recommendations. The configurations allow the manager to create, delete, and update metrics and their levels. For example, the manager can create a metric called "complexity" and add the levels "low, medium, high" for it. The manager also can select the Trello board of the project they are managing where EmoReflex Trello plugin is activated and view project and team details by task or by entire Trello board which is given in Fig. 2. Further, the AI button allows manager to receive recommendations to better manage the team based on the ratings the developers have provided for emotions and metrics, as given in Fig. 3.

3.2 Software Architecture

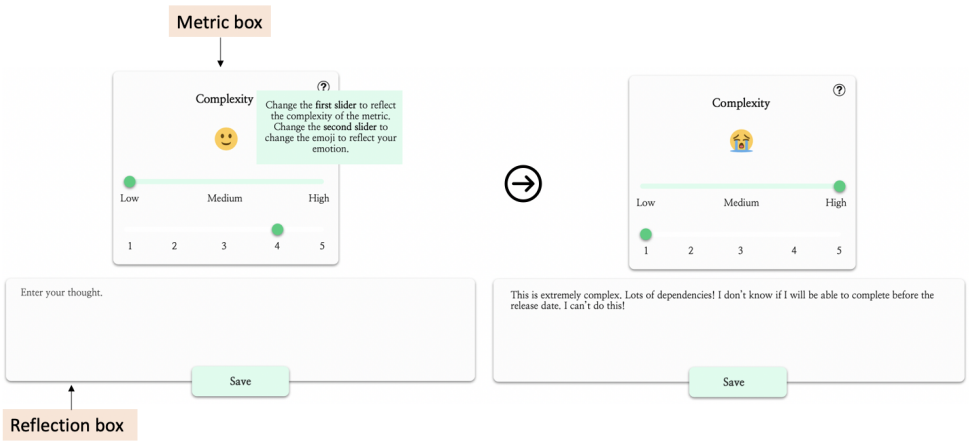


Fig. 1 The box on the top with the emoji is the metric box, and the box on the bottom with text is the reflection box. In this figure, a developer indicates the task in the card has high complexity, feels sad, and enters his thoughts via the reflection box. Hovering over the help icon explains how the sliders work. The save button floats so that the developer does not have to scroll down to save the changes.

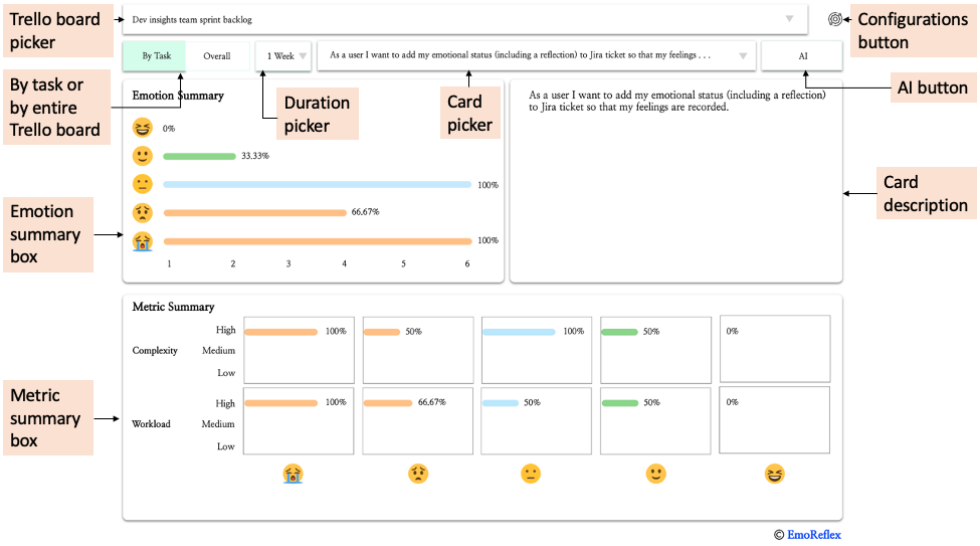


Fig. 2 Once the manager authenticates the Trello account connection with the dashboard, the Trello boards available for their account, the list of cards available, card description, and developers' ratings against the emotions and metrics for each card get loaded to the dashboard. The manager hence picks the Trello board with EmoReflex Trello plugin activation, selects the view, whether it is by task or entire board, and selects the duration for which they need to see the team's emotions and metric ratings. This figure demonstrates a "by task" view. For the "overall" view, the card description is hidden.

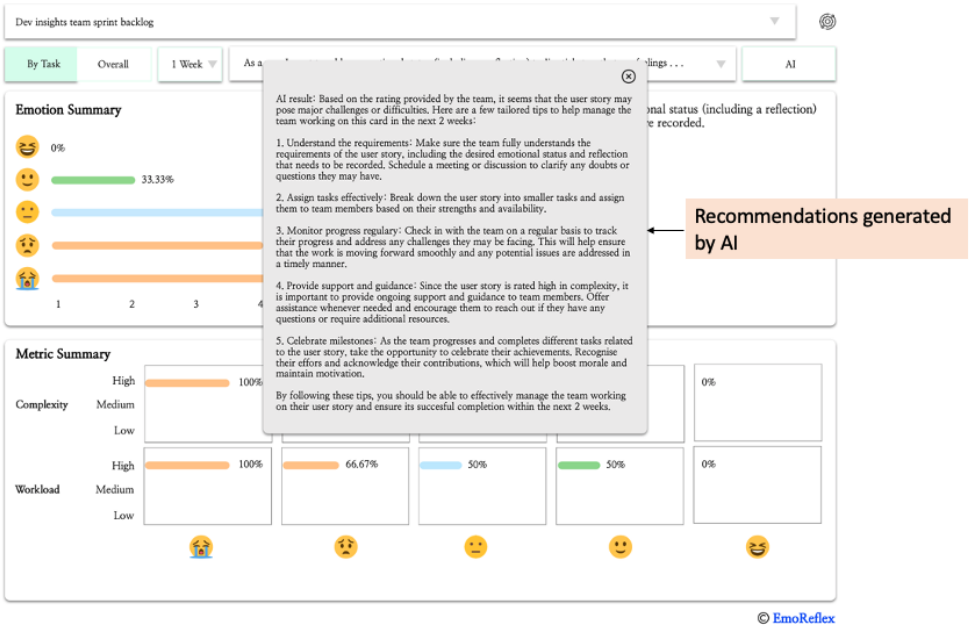


Fig. 3 When the manager clicks on the AI button, the AI generates recommendations for the manager to better handle the team in the next couple of weeks.

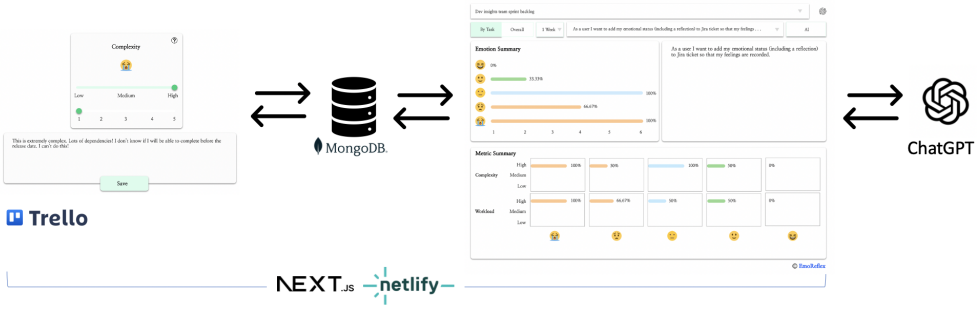


Fig. 4 The software architecture of EmoReflex. The apps are Next.js based and hosted on Netlify. A MongoDB database is used for data storage. The dashboard is also connected to ChatGPT using its API.

3.2.1 Overall structure

The EmoReflex architecture (Fig. 4) is designed to support seamless interaction between its core components—the Trello plugin, the management dashboard, and the AI recommendation engine. At the heart of this setup is a robust server application that handles data processing and integration. The server communicates with a MongoDB database to store and retrieve emotional data metrics as reported by developers

through the Trello plugin. This data is then used to generate insights and recommendations via the AI module, which is powered by ChatGPT API integration. This setup ensures that data flows efficiently between the user interfaces and the storage and processing backends, thereby supporting real-time analytics and feedback mechanisms.

3.2.2 Technology Stack and Implementation Process

EmoReflex's technology stack includes a JavaScript-based Trello plugin and a management dashboard, both developed using Next.js. Next.js was chosen for its server-side rendering capabilities, which enhance performance and SEO. Each component is hosted separately on Netlify, ensuring continuous functionality even during maintenance or updates on other pages. Both the Trello plugin and the dashboard are connected directly to a MongoDB database, allowing real-time updates across all components. The dashboard is integrated with ChatGPT via its API, triggered by click events to minimize unnecessary API calls and reduce server load. The user interface is styled using Tailwind CSS, providing a highly customizable and responsive design. Additionally, the codebase is rigorously tested with Jest, and Continuous Integration and Continuous Deployment (CI/CD) are implemented using GitHub Actions, ensuring high code quality and seamless updates. This is explained further below.

- **Frontend and Backend Development:** Both the Trello plugin and the management dashboard are built using Next.js for its server-side rendering capabilities. Tailwind CSS was used to create a responsive user interface, ensuring compatibility across devices.
- **Database Integration:** A MongoDB database was chosen for its flexible document model, storing emotional data and feedback efficiently. API endpoints were created to manage communication between the database and the front-end components.
- **AI Engine Setup:** The ChatGPT API is connected via server-side logic within the dashboard. To reduce server load, the AI is triggered only upon specific user actions (e.g., a manager clicking the "Generate Recommendations" button).
- **Deployment and Hosting:** Netlify hosts both the Trello plugin and the dashboard, providing automatic builds and updates with minimal downtime. Components are deployed independently, ensuring continuous functionality even during updates to specific modules.
- **CI/CD Pipeline:** GitHub Actions is used to automate testing and deployment. Every commit triggers Jest tests, ensuring the code is reliable and free of critical bugs before deployment.
- **Testing Process:** The implementation was guided by unit testing and integration testing. Beta users provided feedback on usability and performance, which informed iterative improvements before the final release.

3.2.3 Design Principles and Component Breakdown

The implementation of EmoReflex followed single responsibility principle and interface segregation principle. Due to the implementation of the single responsibility principle, each component of EmoReflex is responsible for a single, well-defined functionality, and

due to the implementation of the interface segregation principle, EmoReflex segregates its interfaces based on the specific needs of the components.

Trello Plugin (Frontend Layer): The Trello plugins are embedded using iFrames to seamlessly integrate with the respective platforms without requiring major modifications to the host applications. The user interface design is minimal and intuitive, featuring components such as:

- A rating component that allows users to express their emotions at different stages of an agile project.
- A reflection box for users to provide textual feedback, enabling them to elaborate on their emotions and experiences.

These plugins are developed using the Next.js framework, which allows efficient server-side rendering (SSR) and easy deployment on Netlify. Authentication mechanisms are implemented using NextAuth, ensuring secure and seamless login for users across the platforms. This also helps manage user sessions and permissions.

Backend Server (API Layer): The backend server is built using Next.js with TypeScript, acting as the central hub for processing requests, storing data, and handling AI recommendations. It provides a robust API layer that manages the flow of data between the frontend plugins, the AI engine, and the database. RESTful APIs are used for communication, ensuring standardized data exchange between components. The backend server processes incoming requests, validates data, and sends responses to the plugins, ensuring a smooth user experience.

Dashboard (Frontend and Backend Layer): The dashboard provides a high-level overview of emotions and feedback collected from the Trello and Jira plugins, giving teams and project managers insights into team sentiment and engagement. The Dashboard Frontend is built using Next.js with SSR, ensuring that data is preloaded on request to enhance the user experience with fast page rendering. The Dashboard Backend extracts and processes data from MongoDB to display insights such as team emotion trends over time, individual reflections and recommendations, and comparative analysis between project stages or sprints. The dashboard's fron end communicates with the backend via API calls to fetch processed data and display it in a visually engaging format.

3.2.4 Communication and Data Flow

The following explains the communication and data flow (request/response flow) of EmoReflex.

- **User Input:** A user submits an emotion rating through the Trello or Jira plugin.
- **Data Transmission:** The plugin sends the user's input to the backend server using API calls.
- **Data Processing and Storage:** The backend processes the input and stores it in the MongoDB database using Prisma.
- **Generating Insights:** The backend sends emotional data from the database to the dashboard to display and generate insights using ChatGPT with a structured prompt.

- **Change Plugin Configuration:** Managers can modify what emotions and metrics are tracked for a specific project from the Dashboard.

3.2.5 Deployment Architecture

Hosting on Netlify: Netlify is chosen as the hosting platform for its simplicity, seamless integration with Next.js, and robust CI/CD support. It automates the deployment process and ensures that every push to the GitHub repository triggers a build and deployment, allowing for continuous delivery.
CI/CD Pipelines: GitHub Workflows are employed to automate the CI/CD pipeline. Every change pushed to the repository triggers automated tests, builds, and deployments. This ensures that only verified code reaches production, maintaining system stability and reducing the chance of errors.

3.2.6 Usability

The design of EmoReflex is highly user-centric, ensuring that user interactions are as simple and intuitive as possible with multiple iterations of prototype development at different levels. We started with an initial design stage where we came up with low-fidelity wireframes where first and second authors (acting as product manager and client) gave feedback on, then in the next phase of beta testing, we built and deployed a functional prototype for 2-3 weeks within a controlled environment which was tested by the first author and the rest of the team. We then had ongoing feedback loops to allow for further refinement, ensuring the platform evolves according to user needs. The action to rate emotions has been streamlined to minimize interruptions to the user's (developer's) tasks, allowing them to quickly and easily record their emotional state. The dashboard is designed to display important data and insights by default, providing users (managers) with immediate access to critical information such as the *overall* emotion and metric summary and of the team, while still allowing for more detailed exploration, such as team's emotion summary and metric summary *by task*, if desired. Additionally, the design follows WCAG 2.1 guidelines to ensure accessibility, and all colors have been carefully selected to be inclusive of color-blind users, ensuring that everyone can use the platform seamlessly and effectively.

3.2.7 Maintenance and Extensibility

The modular architecture ensures each component operates independently while staying connected through standardized APIs. This modularity simplifies maintenance and allows for future extensions, such as integrating EmoReflex with Jira or other project management tools. New features can be introduced without disrupting existing functionality due to the CI/CD pipeline. Automated testing with Jest ensures that updates meet quality standards, while detailed documentation helps onboard new developers quickly. Although the platform is designed as a prototype, these practices ensure it remains reliable and adaptable to evolving requirements.

4 User Evaluation

We evaluated the two components of EmoReflex from the perspectives of the two user groups. i.e., professional software developers working in team contexts evaluated the Trello plugin and managers evaluated the dashboard. We wanted to assess the mental workload of the two user groups when using EmoReflex in different use cases. Hence, we used the NASA Task Load Index for each use case in the user evaluation We also wanted to assess the usability of EmoReflex as a whole. Hence, we used the system usability scale. Further, we collected qualitative feedback on positive/negative experiences and potential experiences from the two user groups which are necessary for future development of EmoReflex.

4.1 Approach

Study Protocol. We took a form based approach to gather feedback from the participants[1]. Since we had two user groups, we used two different forms with use cases (Figures 5, 6, 7) distinctive to the two groups. Each form had an overview description of the tool, questions on participant demographics, followed by quantitative and qualitative questions relevant to each use case. The use cases accompanied videos of the use cases in action as well. Therefore, participants watched the videos first and answered its questions next. The manager feedback form had 47 questions and the developer feedback form had 35 questions including one for the participants to enter their Prolific ID. The reason why the number of questions differed for the two user groups was as the manager had more use cases than the developer. We used NASA Task Load Index for each of the use cases and the system usability scale (SUS) test for the entire tool for both user groups. Both NASA Task Load Index and SUS were quantitative questions. We further had open-ended questions for participants to share their positive/negative experiences and their opinions on potential improvements of EmoReflex. The feedback forms are available online[2].

Participant Recruitment. We recruited participants (20 developers and 20 managers) using the participant recruitment platform Prolific. We have had positive experience using Prolific in the past. Therefore, we decided to use it in this study as well. We used pre-screening questions of working as an individual contributor and as a manager in IT. We paid 4.50 Sterling Pounds for each participant for their time.

Data Collection. The data were collected on the two different Google forms which encompasses both quantitative and qualitative questions. An average time of 30 minutes were spent by the participants to fill out the form including watching the videos.

Data Analysis. We used Microsoft Excel to analyse the quantitative data. We used ChatGPT as a co-analysis tool to support the qualitative data analysis. Since the use of ChatGPT for qualitative analysis in scientific research is relatively new, we explain our procedure further here. First, we went through each qualitative data point to segment them as positive, negative and potential improvements. Then, to ChatGPT, we fed each data segment and asked to conduct inductive thematic analysis.

[1] Approved by Monash Human Research Ethics Committee Approval Number: 41279
[2] https://github.com/kashumi-m/ReplicationPackageEmoReflexUserEvaluation

Scenario A - Metric box:

1. Click on EmoReflex on the Trello card.
2. Use the first slider to rate how you find the metric related to the task on the card. For the example below, depending on the task given on the Trello card, you can indicate whether the metric "complexity" is "low/medium/high" for you.
3. Click "Save" to save your ratings. Your responses along with your teammates' responses will get aggregated together and will be displayed on manager's dashboard. Your identity will not be revealed to the manager.

Scenario B - Reflection box:

1. Type what you feel about the task on the Trello card.
2. Click "Save" to save how your text. This will not be sent to your manager. You will be able to get a record of what you entered as a PDF later. This is not available at the moment. Only "Save" option is available.

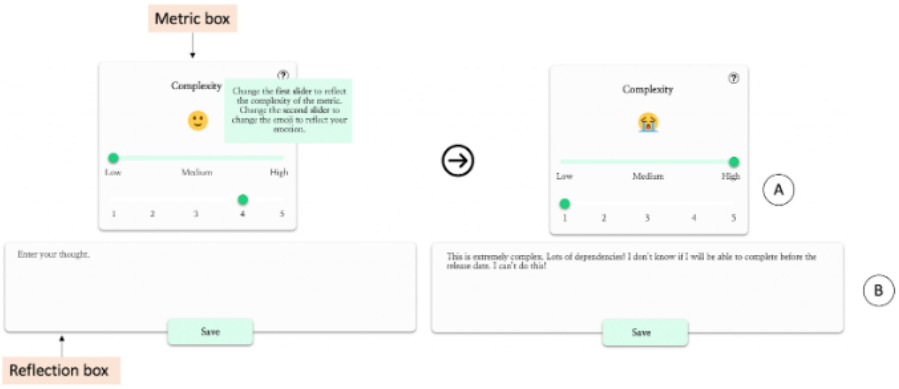


Fig. 5 This is how the use cases were shown on the developer form. Use cases A an B as in this paper reflect the scenariso A and B as in the developer form. i.e., as in this figure. For further information/questions refer the replication package.

Then it came up with themes for the data. We further asked it to map each of the data points against the themes it came up with. This allowed us to quality control the process. Once the thematic analysis was done, we asked it to further group the themes to high level themes. At each step, we thoroughly went through its analysis so that we maintained the reliability, accuracy, and rigour of the analysis. We fixed its analysis a couple of times when we could not agree with its analysis. Overall, we had a positive experience analysing the qualitative data using ChatGPT given a high agreement with its analysis. We did not calculate the inter-rater agreement as we used ChatGPT as a co-analysis tool but not as a separate analyser.

The first author, who is a well experienced data analyst in academia and industry, analysed all the data. As ChatGPT was involved in the qualitative analysis process, the data analysis is not completely biased towards the analysis of the human analyser. The responses by ChatGPT for the prompts on analysis followed with a thorough quality

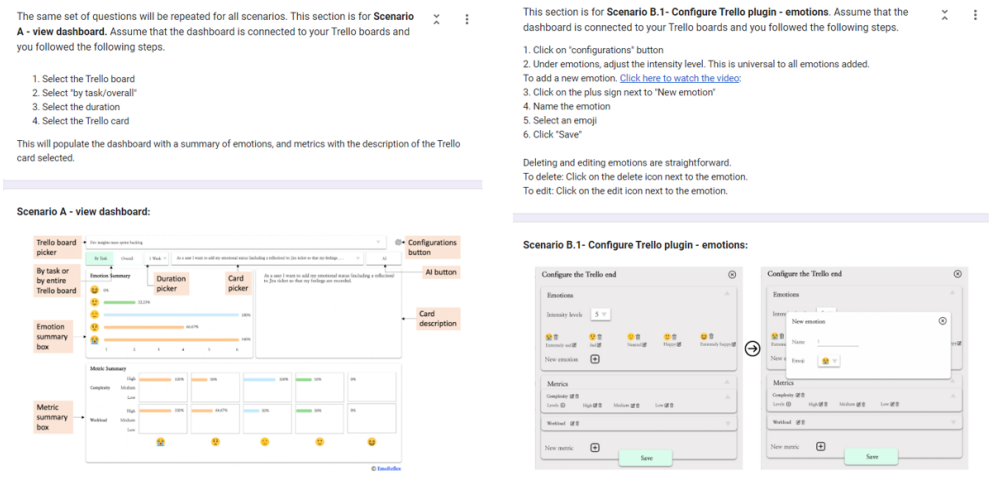


Fig. 6 This is how the use case A and B.1 were shown on the manager form. Use cases A an B.1 as in this paper reflect the scenarios A and B.1 as in the manager form. i.e., as in this figure. For further information/questions and better quality images refer the replication package.

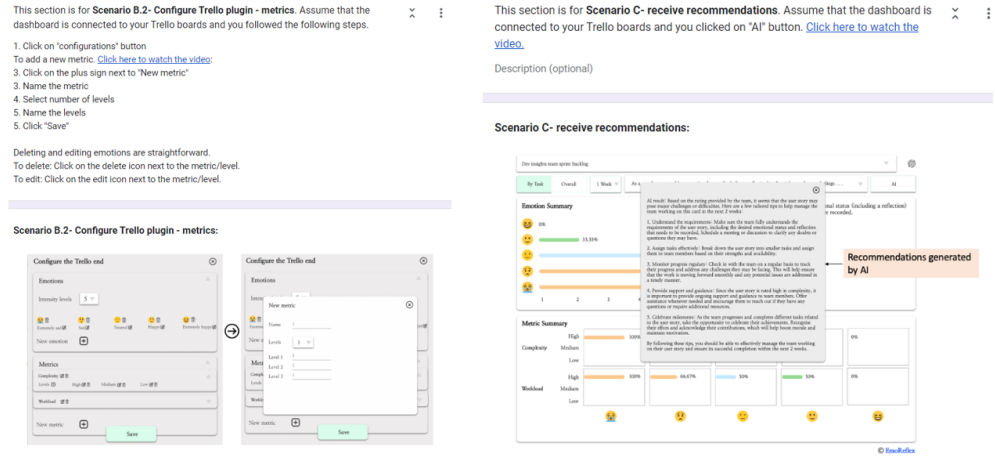


Fig. 7 This is how the use cases B.2 and C were shown on the manager form. Use cases B.2 and C as in this paper reflect the scenarios B.2 and C as in the manager form. i.e., as in this figure. For further information/questions and better quality images refer the replication package.

check. For example, ChatGPT was asked to provide the mapping of the data points provided against the themes it created. These data points were then rechecked to see if ChatGPT was hallucinating and to see if all data points were mapped (Example in Fig. 8). When ChatGPT was used to analyse the first set of data points, the first author checked each and every analysis, and corrected if necessary – such as renaming the themes (ChatGPT was asked to provide explanations (Example in Fig. 9) for the

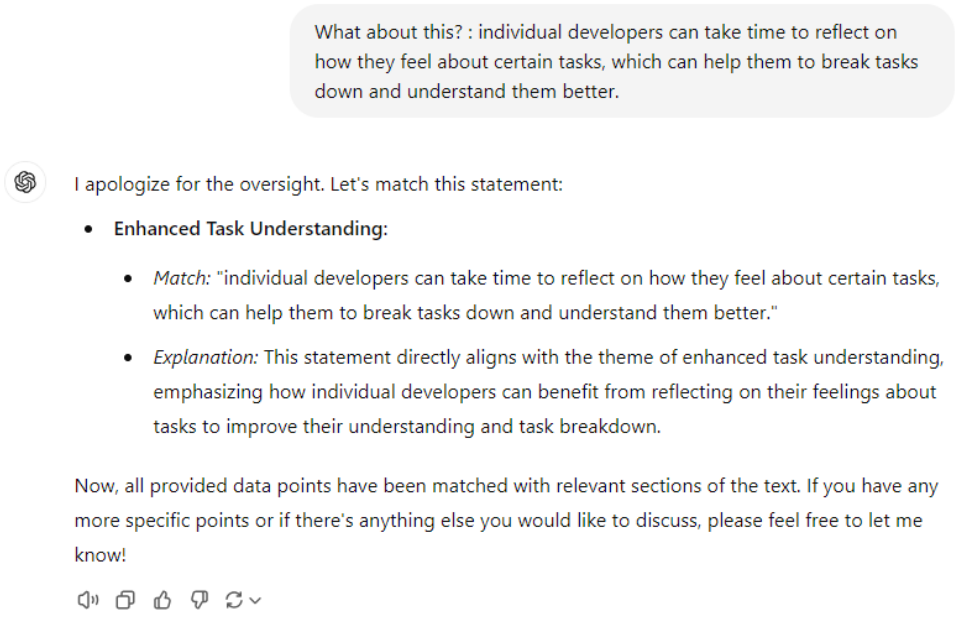


Fig. 8 The data points were rechecked to see if ChatGPT was hallucinating and to see if all data points were mapped.

themes to understand its analysis too) and recategorising, and moved to the next set of data. This was repeated for the entire dataset. The findings were shared with the second author with the working draft of the paper, which was done simultaneously. The findings were reviewed by the second author who provided feedback such as on some of the themes that could be better merged cohesively into high-level themes and better nomenclature. The findings were revised such that both the first and second authors agreed and finalised. The quotes reported in the paper reflect the accuracy of the analysis.

4.2 Evaluation Results

4.2.1 Developer Evaluation Results

Developers: Participant Information.

Demographic information of the developers who participated in our study is given in Table 1. The majority of the developers participated were in the age group of 26-30 years (n=9), the majority were men (n=17), and the majority lived in Australia (n=5). 6 participants mentioned that they were neurodiversed (having a concentration and/or memory disorder, mood or emotional disorder, autism/autism spectrum disorder, learning differences) while 2 of them said they were undiagnosed. The developers had a median of 2.50, a mean of 3.78, and a standard deviation of 3.83 years of professional developer experience. The majority of the participants worked in the IT

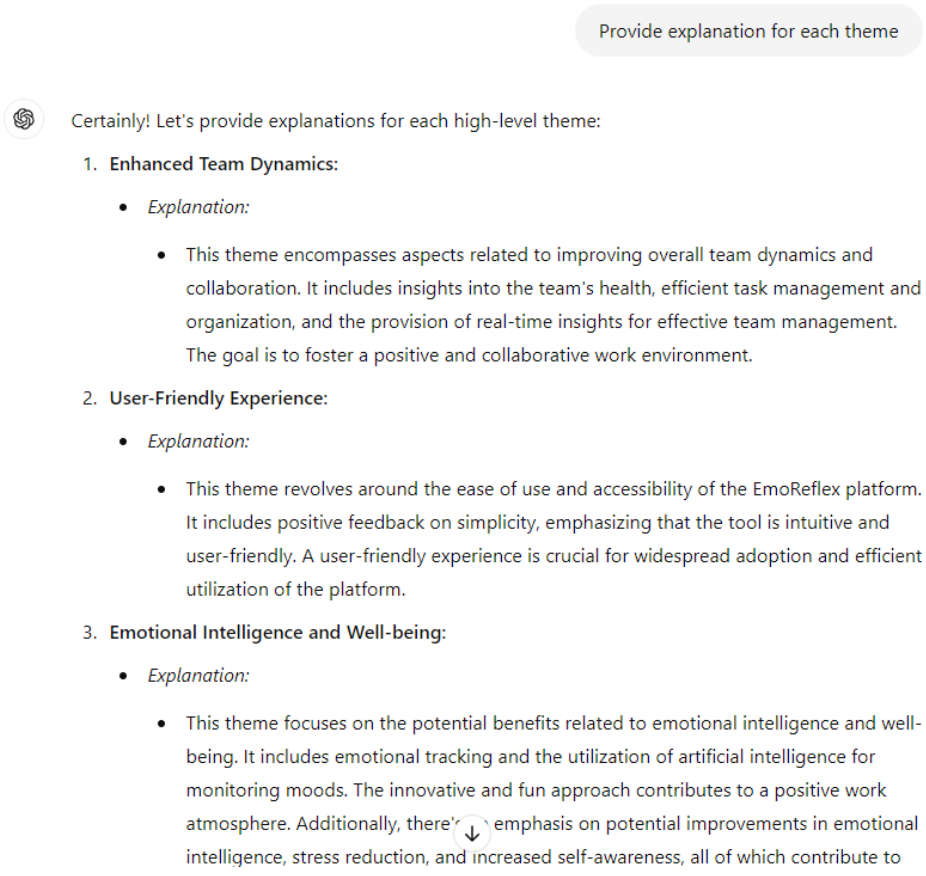
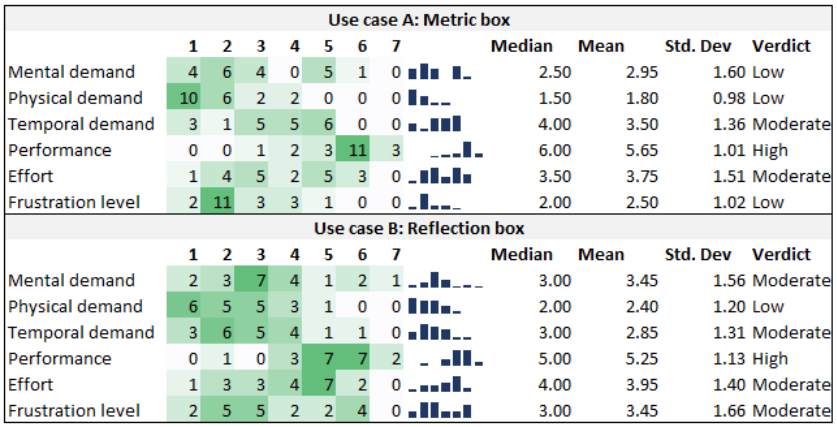


Fig. 9 ChatGPT was asked to provide explanations for the themes it created.

industry (n=15) and majority worked in organisations with employees 11-50 (n=15). Participants' team size's median was 8, means was 11.90 with a standard deviation of 10.73.

Table 1 Developer information. 20 developers participated in the study (Example of autism/an autism spectrum disorder: Aperger's, Example of a concentration and/or memory disorder: ADHD; Examples of emotional disorder: depression, bipolar disorder, Examples of learning differences: Dyslexic, Dyslexia).

ID	Age	Gender	Country	Neurodiversity	Professional developer experience (years)	Industry	Organisation size	Team size
DP1	26-30	Man	Australia	*ADHD - Undiagnosed*	4	Healthcare	11-50	25
DP2	26-30	Man	Australia	-	1	IT	More than 10000	5
DP3	20-25	Woman	Australia	Concentration and/or memory disorder	4	IT	501-1000	50
DP4	26-30	Man	Netherlands	Autism/an autism spectrum disorder	6	IT	11-50	4
DP5	26-30	Man	Australia	-	4	IT	101-500	10
DP6	26-30	Man	Mexico	Mood or emotional disorder	2	IT	5001-10000	22
DP7	20-25	Man	Australia	-	2	IT	11-50	5
DP8	20-25	Man	Netherlands	Learning differences	5	IT	11-50	20
DP9	36-40	Woman	Netherlands	Concentration and/or memory disorder	11	Medical	5001-10000	17
DP10	31-35	Man	Germany	-	1	Transport	11-50	5
DP11	20-25	Man	United Kingdom	-	1.5	IT	11-50	5
DP12	26-30	Man	Netherlands	-	3	Manufacturing	1001-5000	9
DP13	26-30	Man	Portugal	Learning differences	1	Manufacturing	11-50	4
DP14	20-25	Man	Poland	Mood or emotional disorder	2	IT	11-50	4
DP15	20-25	Man	Portugal	-	1	IT	More than 10000	12
DP16	41-45	Man	Canada	-	17	IT	1001-5000	6
DP17	31-35	Man	Germany	*"While not diagnosed, I get distracted easily. "*	4	IT	1001-5000	12
DP18	31-35	Man	Portugal	-	1.5	IT	51-100	6
DP19	26-30	Man	Canada	-	1	IT	101-500	7
DP20	26-30	Woman	France	-	3.5	IT	More than 10000	10



Use case A: Metric box											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	4	6	4	0	5	1	0	2.50	2.95	1.60	Low
Physical demand	10	6	2	2	0	0	0	1.50	1.80	0.98	Low
Temporal demand	3	1	5	5	6	0	0	4.00	3.50	1.36	Moderate
Performance	0	0	1	2	3	11	3	6.00	5.65	1.01	High
Effort	1	4	5	2	5	3	0	3.50	3.75	1.51	Moderate
Frustration level	2	11	3	3	1	0	0	2.00	2.50	1.02	Low
Use case B: Reflection box											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	2	3	7	4	1	2	1	3.00	3.45	1.56	Moderate
Physical demand	6	5	5	3	1	0	0	2.00	2.40	1.20	Low
Temporal demand	3	6	5	4	1	1	0	3.00	2.85	1.31	Moderate
Performance	0	1	0	3	7	7	2	5.00	5.25	1.13	High
Effort	1	3	3	4	7	2	0	4.00	3.95	1.40	Moderate
Frustration level	2	5	5	2	2	4	0	3.00	3.45	1.66	Moderate

Fig. 10 Mental work load test results of developers. The mental demand was low for use case A (metric box) but moderate for use case B (reflection box). The physical demand was low and temporal demand was moderate in both cases. The performance was high in both cases and effort was moderate in both cases. The frustration level was low for use case A but moderate for use case B. (1-7: likert scale; Std. Dev: standard deviation; the number of participants rated for each item in the scale against the likert scale are given).

Developers: Mental Workload Test Results. The NASA Task Load Index test results for developers are given in Fig. 10. The mental demand was low for use case A (metric box) but moderate for use case B (reflection box). The physical demand was low and temporal demand was moderate in both cases. The performance was high in both cases and effort was moderate in both cases. The frustration level was low for use case A but moderate for use case B.

Developers: System Usability Test Results. The SUS results of developers are given in Fig. 11. We followed the scoring system to calculate the SUS score of the test. It results in a score of 71.88 which is grade B in industry standards (a system with a SUS score of 68 is considered the average). That means, the system is acceptable by the users (NPS (Net Promoter Score) level: "passive") but can be improved (equivalent to NPS level "promoter". i.e., the users promote/recommend the system to others). Note that the system is at the prototype level, meaning there is room for improvement before releasing to the market.

Developers: Qualitative Feedback. The developers shared positives, negatives, and potential improvements for EmoReflex. The summary of feedback is given in Fig. 12 and detailed version is available in Appendix A.

👍 *Positive feedback: User-friendly experience.* As per [DP5], "it is quick and easy", and as per [DP14], "easy to use, minimal & good design, anonymous". And if EmoReflex is used, "I could see an improvement in communication; especially if people are working in a team remotely it might be hard to gauge their feelings regarding project tasks" [DP20]. "This tool would be good to kinda test the waters of what my team is feeling on any given tasks. This could also be helpful for more shy members of the team that might have trouble expressing their feelings" [DP19]. Further, [DP20] said, "easy to use and more playful". On transparent communication, as said by [DP17], "I like the anonymous submission of comments and the use of the complexity emojis.

SUS Score: 71.88 (Grade: B)	1	2	3	4	5	Median	Mean	Std Dev
I think I would like to use this system frequently.	2	5	5	8	0	3	2.95	1.02
I found the system unnecessarily complex.	9	6	2	3	0	2	1.95	1.07
I thought the system was easy to use.	0	1	3	11	5	4	4.00	0.77
I think that I would need the support of a technical person to be able to use this system.	13	4	3	0	0	1	1.50	0.74
I found the various functions in this system were well integrated.	0	1	6	10	3	4	3.75	0.77
I thought there was too much inconsistency in this system.	5	8	5	2	0	2	2.20	0.93
I would imagine that most people would learn to use this system very quickly.	0	1	3	8	8	4	4.15	0.85
I found the system very cumbersome to use.	8	4	3	4	1	2	2.30	1.31
I felt very confident using the system.	1	2	6	10	1	4	3.40	0.92
I needed to learn a lot of things before I could get going with this system.	11	9	0	0	0	1	1.45	0.50

Fig. 11 System usability scale results of developers. The SUS score resulted was 71.88 which is grade B (1-5: likert scale; Std. Dev: standard deviation; the number of participants rated for each item in the scale against the likert scale are given).

Developer feedback	Participants	
Positives		
User-friendly experience	6	30%
Effective task management	5	25%
Managerial decision support and collaboration	3	15%
Professional development support	1	5%
Customisation and metrics	1	5%
Negatives		
Communication and feedback preferences	4	20%
Complexity and tool dependency	2	10%
Other concerns and resistance	6	30%
Potential improvements		
Confidence and transparency	1	5%
Flexibility and optional features	2	10%
Functionality and metrics	3	15%
Satisfaction and purpose	6	30%
User experience and interface	7	35%

Fig. 12 The developers shared positives, negatives, and potential improvements for EmoReflex.

I could see that this would remove a lot of discrepancies during sprint planning and refinement".

👍 *Positive feedback: Effective task management.* Enhanced task understanding was a positive feedback we received from the participants. [DP2] mentioned "individual developers can take time to reflect on how they feel about certain tasks, which can help them to break tasks down and understand them better" and [DP1] mentioned "I think the pros include assigning a task difficulty" where "difficulty" could be any pre-defined metric.

[DP11] commented on the data utilisation. They said "I suppose the data collected could be out to good use potentially".

Further, [DP3] said "developers get to express an emotion related to the task, which might indicate to managers which are more desirable tasks for certain people" indicating the improved task assignment.

EmoReflex supports strategic rating and retrospective facilitation according to [DP9]. They said "retrospectives will be easier, helping recollect task issues. Helps rate task refinement, I.e. if complexity was too high, should there have been smaller tasks? If poor emotions, should more internal/external support have been provided? Also keeps track of good practices with whatever went well".

👍 *Positive feedback: Managerial decision support and collaboration.* Informed managerial decision-making as a pro of EmoReflex as commented by [DP2] where "can help managers obtain a broader view of how their delegates feel about work tasks, and they could use these insights to implement processes to improve attitudes to work". It also helps "To quickly showcase to my manager if I'm suited for this task" [DP16], and as per [DP2] "If everyone spent time beforehand using this tool, our project manager might agree with our assessments of tickets :)" which improves project management agreement.

👍 *Positive feedback: Professional development support.* "It could be a nice way to quickly determine the professionalism changes needed, which groups need more training, overall any strategy ratings" [DP8] which explains the support to professional development.

👍 *Positive feedback: Customisation and metrics.* [DP18] commented briefly that one of EmoReflex's pros is "customisable metrics and emotions".

👎 *Negative feedback: Communication and feedback preferences.* A couple of participants showed desire for more information and clarity. They expressed a desire for clearer metrics, "Need more numbers maybe from 1 to 10" [DP13], and additional information for handling more complex tasks "Not enough info for complex tasks" [DP14]. However, these can be configured at the manager-end. [DP3] raised a concern on anonymity and openness. They were concerned "if the system was not anonymous". [DP4] showed a preference for traditional feedback, "I feel like there are better traditional ways of handling developer feedback if their struggling with a task, such as simply asking a colleague".

👎 *Negative feedback: Complexity and tool dependency.* [DP16] said, "It's a bit complicated" and [DP18] said EmoReflex is "Limited reflection options, potential for bias in metrics, dependency on Trello".

👎 *Negative feedback: Other concerns and resistance.* [DP11] expressed annoyance and lack of consistent usage of EmoReflex. They said "But I don't think people would consistently use the system if they didn't have to. I think I'd find it annoying". [DP2] raised concens about time and engagement. They said "Employees might get hung up on trying to reflect on every single task and therefore spend a disproportionate amount of time using the tool as opposed to actually working on their tasks. They might also not engage with the tool at all, rendering it useless, or providing a skewed overview of responses if only one or two people respond to it". [DP12] said "But I'm worried about privacy", sharing their opinion on privacy concerns. However, they did not further comment on this. [DP8] said " Perhaps if it's a mandatory option then it would come with skewed results", depicting a resistance to mandatory usage. [DP1] had a skepticism about emotional aspect. They said " but I dont believe the emotional aspect is required. I think its unnecessary additional work. I think its unnecessary extra work that has no benefit". From [DP10]'s perspective, they have an uncertainity

about benefits. As shared, "To be honest, I don't know what benefits it will have in business life".

Potential improvement: Confidence and transparency. [DP9] commented on confidence and transparency as a potential improvement. They said, "Ability to view past manager dashboards, to give more confidence that private comments are not visible, and to help understand team better, perhaps as a part of retrospective sessions" could improve the tool more.

Potential improvement: Flexibility and optional features. A couple of participants commented on flexibility and optional features. [DP3] said, "Might be good to have optional sliders, so in cases where the complexity or emotion is hard to determine they do not need to be entered. This is also assuming that feedback/ reflection is also optional", and [DP18] said, ""Allow users to edit and format their reflections for a more nuanced expression. ex: tagging, categorization".

Potential improvement: Functionality and metrics. Some participants commented on improving the functionality and metrics of the tool. "Broaded the range of the task difficulty to maybe 1-10. No need for emotion component at all. Just a comlexity scale with more variance" [DP1], "More metrics and analysis tools for the dashboard" [DP6], and "Add more emotions. More emotional range, and maybe give context to the number scale. Like what EXACTLY is a "5", is it ease of doing a task, is it complexity? I really don't know. Having the Help button there is nice, maybe include a little blurb defining what the intended scale is". The variety of feedback explains the contradiction of some liking the emotional component and some disliking it, which we believe can be solved by having an ON/OFF switch for it.

Potential improvement: Satisfaction and purpose. Some participants showed an overall satisfaction and functionality. "Everything is working fine" [DP5], "Not sure, it seems to fulfil its purpose OK" [DP11], "Time saver, easy to manage team's opinion about specific tasks" [DP14], "I think it serves well enough for what its purpose is" [DP15].

[DP4]'s feedback indicated potential pragmatic acceptance. They said "For what it is, I think this is fine. If I'd be forced to use it, then atleast I can get through the system quickly enough that it wouldn't be much of a bother". [DP12] could not see any improvements, but had an uncertainty and resistance towards the tool. They said "I can't think of any improvements. If I have a choice, I wouldn't use this tool at work".

Potential improvement: User experience and interface. Some participants said there could be improvements in the user interface and ease of use. They said, "An improvement to the UI and ease of use" [DP2], "Better, cleaner UI (esp transition anims)" [DP8], "It can be more user friendly" [DP16], and "Everything looked good, maybe polish the UI up a bit (different set of emojis for example)" [DP17].

Further, the participants commented on improving aesthetic and emotional component. They shared, "I think the emotion popup UI needs more work to be aesthetic, hence, creates a better user experience" [DP7], "More numbers to 1 to 10 and maybe take the emojis out" [DP13], and "Also, a tag or category system might be a cool feature when rating. For example, a crying emoji could be tagged with multiple predefined options" [DP17].

4.2.2 Manager Evaluation Results

Managers: Participant Information.

The details of software engineering managers participated in our study are given in Table 2. The majority of the managers participated in our study were in the 31-25 age bracket (n=7), majority were men (n=15), and majority were from Australia (n=5). 5 participants were neurodiverse – 3 had a concentration and/or memory disorder, 1 experienced learning differences, and 1 had anxiety disorder. The participants had a median of 3, a mean of 4.24 with a standard deviation of 3.30 years of software engineering management experience. Most of the partipants worked in the IT industry (n=12) and most worked in organisations which had more than 10000 employees (n=4). The participants' team size had a median 10, a mean of 18.24, and a standard deviation of 27.37. One participant had a team size of 125 members which is quite unusual for a development team. However, as the participant was from an organisation with more than 10000 employees, we assume that this could be a very large team consisting of multiple small teams.

Table 2 Manager information. 20 software engineering managers participated in the study (Example of concentration and/or memory disorders: ADHD; Examples of learning differences: Dyslexic, Dyslexia).

ID	Age	Gender	Country	Neurodiversity	SE Manager experience (years)	Industry	Organisation size	Team size
MP1	41-45	Woman	Australia	-	9	Telecom	More than 10000	7
MP2	36-40	Man	Australia	-	2	Finance and Banking	5001-10000	4
MP3	31-35	Man	Germany	-	3	IT	1001-5000	25
MP4	26-30	Man	Australia	-	2	Finance and Banking	51-100	4
MP5	26-30	Man	Portugal	Concentration and/or memory disorder	2.5	IT	More than 10000	10
MP6	31-35	Woman	Australia	-	3	Healthcare	5001-10000	4
MP7	51 and above	Man	Australia	-	3	IT	51-100	11
MP8	20-25	Man	Poland	-	1.5	IT	11-50	7
MP9	41-45	Man	Poland	Learning differences	10	IT	501-1000	20
MP10	31-35	Woman	Poland	-	4	IT	More than 10000	125
MP11	31-35	Man	Spain	-	3	IT	101-500	10
MP12	31-35	Man	Greece	-	3	IT	11-50	5
MP13	36-40	Man	Mexico	-	2	Manufacturing	5001-10000	10
MP14	26-30	Man	Portugal	Concentration and/or memory disorder	2.5	IT	More than 10000	10
MP15	31-35	Woman	Australia	-	4	IT	1001-5000	4
MP16	36-40	Woman	Germany	-	3.5	Finance and Banking	101-500	14
MP17	31-35	Man	Germany	Anxiety disorder	5	IT	101-500	25
MP18	46-50	Man	France	-	15	IT	501-1000	60
MP19	20-25	Man	Portugal	Concentration and/or memory disorder	2	Manufacturing	11-50	10
MP20	36-40	Man	Mexico	-	6	Manufacturing	1001-5000	8

Managers: Mental Work Load Test Results. The NASA Task Load Index test results for the 3 use cases for managers (A: view dashboard, B.1: configure Trello plugin - emotion, B.2: configure Trello plugin - metrics, C: receive AI recommendations) are given in Fig. 13. In all cases, the mental demand, physical demand, temporal demand, effort, and frustration were low. The performance was high.

Managers: System Usability Test Results. The SUS results of the managers are given in Fig. 14. Similar to the developer user study, we used the SUS scoring system to calculate the SUS score of our test. The SUS score resulted for managers was 74.63 which is also grade B in industry standards.

Managers: Qualitative Feedback. The managers shared positives, negatives, and potential improvements for EmoReflex. The summary of feedback is given in Fig. 15 and detailed version is available in Appendix B.

👍 *Positive feedback: Enhanced team dynamics.* The managers' insights on enhanced team dynamics highlighted EmoReflex's ability to provide team insights and at-a-glance view, support task management and organisation, and provide real-time insights for team management.

The managers feedback depicted that they found EmoReflex "provides insights of the team's health" [MP1] and "at-a-glance view is great" [MP2]. As shared by [MP14], it is – "a better perspective of how the workers are doing and perhaps I might notice some patterns I haven't noticed before (like an employee always being sad or stressed on the same day of the week)". They also stated EmoReflex "might help the management team to get there work well organized and a good way of ensuring and assigning the tasks to right people" [MP6]. Further, it "enhances team management by providing real-time insights into team emotions and performance, enabling tailored support and improved decision-making" [MP18].

👍 *Positive feedback: User-friendly experience.* The managers' feedback on user-friendly experience brought out that EmoReflex is user-friendly and intuitive, and they further provided positive feedback on simplicity.

The participants found that EmoReflex is "intuitive and easy to configure for potential projects " [MP7] and it "can streamline task management within Trello, making it easier for teams to collaborate, track progress, and meet deadlines" [MP16]. They also said "it is a relatively simple tool to use. On the (other) hand, it is a relatively simple aspect to learn to use" [MP20].

👍 *Positive feedback: Emotional intelligence and well-being.* The participants identified that EmoReflex supports emotion tracking and AI utilisation, supports enhanced emotional intelligence and well-being, and it's an innovative and fun approach.

[MP11] shared "being able to monitor my moods and that of my team, the use of artificial intelligence opens up a whole world of possibilities". Also, "EmoReflex could help employees develop their emotional intelligence (EQ), which can lead to better communication, teamwork, and conflict resolution. This could create a more positive and productive work environment for everyone" [MP16]. Further, they found "this idea is quite innovative and it would really help managers to ease there tasks and its fun way to do it with emotions which is kinda new!" [MP15].

Use case A: View dashboard											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	6	3	4	5	1	0	1	3.00	2.81	1.60	Low
Physical demand	10	2	2	3	1	2	0	1.50	2.40	1.75	Low
Temporal demand	6	5	6	1	1	1	0	2.00	2.43	1.36	Low
Performance	0	0	3	1	4	9	3	6.00	5.43	1.24	High
Effort	6	5	3	5	0	1	0	2.00	2.52	1.40	Low
Frustration level	9	7	2	1	0	0	1	2.00	2.00	1.41	Low
Use case B.1: Configure Trello plugin - emotion											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	7	3	7	2	1	0	0	2.50	2.36	1.19	Low
Physical demand	11	2	5	0	1	1	0	1.00	2.00	1.43	Low
Temporal demand	6	6	6	0	1	1	0	2.00	2.33	1.31	Low
Performance	0	1	3	1	2	8	5	6.00	5.43	1.53	High
Effort	7	4	6	2	1	0	0	2.00	2.29	1.19	Low
Frustration level	11	4	4	0	1	0	0	1.00	1.76	1.08	Low
Use case B.2: Configure Trello plugin - metrics											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	6	6	4	3	0	1	0	2.00	2.38	1.32	Low
Physical demand	9	7	1	2	0	1	0	2.00	2.00	1.30	Low
Temporal demand	5	11	1	2	1	0	0	2.00	2.14	1.06	Low
Performance	0	2	1	0	5	6	6	6.00	5.52	1.53	High
Effort	8	3	6	3	0	0	0	2.00	2.19	1.12	Low
Frustration level	13	6	0	0	1	0	0	1.00	1.48	0.92	Low
Use case C: Receive AI recommendations											
	1	2	3	4	5	6	7	Median	Mean	Std. Dev	Verdict
Mental demand	5	7	1	6	1	0	0	2.00	2.52	1.28	Low
Physical demand	10	3	4	2	1	0	0	1.50	2.02	1.24	Low
Temporal demand	8	6	3	0	2	1	0	2.00	2.24	1.48	Low
Performance	0	1	1	3	2	8	5	6.00	5.52	1.40	High
Effort	6	5	4	4	1	0	0	2.00	2.43	1.24	Low
Frustration level	11	4	2	3	0	0	0	1.00	1.81	1.11	Low

Fig. 13 Mental work load test results of managers. In all cases, the mental demand, physical demand, temporal demand, effort, and frustration were low. The performance was high (1-7: likert scale; Std. Dev: standard deviation; the number of participants rated for each item in the scale against the likert scale are given).

SUS Score: 74.63 (Grade: B)	1	2	3	4	5	Median	Mean	Std. Dev
I think I would like to use this system frequently.	1	3	5	7	4	4.00	3.52	1.12
I found the system unnecessarily complex.	8	9	2	1	0	2.00	1.81	0.81
I thought the system was easy to use.	0	1	3	10	6	4.00	4.05	0.80
I think that I would need the support of a technical person to be able to use this system.	11	3	2	3	1	1.00	1.95	1.30
I found the various functions in this system were well integrated.	2	0	2	12	4	4.00	3.81	1.08
I thought there was too much inconsistency in this system.	10	8	2	0	0	1.50	1.60	0.66
I would imagine that most people would learn to use this system very quickly.	1	0	2	9	8	4.00	4.14	0.96
I found the system very cumbersome to use.	8	7	2	1	2	2.00	2.10	1.26
I felt very confident using the system.	0	1	3	11	5	4.00	4.00	0.77
I needed to learn a lot of things before I could get going with this system.	6	9	1	4	0	2.00	2.14	1.06

Fig. 14 System usability scale results of managers. The SUS score resulted was 74.63 which is grade B (1-5: likert scale; Std. Dev: standard deviation; the number of participants rated for each item in the scale against the likert scale are given).

Manager feedback	Participants	
Positives		
Enhanced team dynamics	8	40%
User-friendly experience	6	30%
Emotional intelligence and well-being	5	25%
Negatives		
User interaction and experience	5	25%
Other concerns and resistance	9	45%
Potential improvements		
Confidence and transparency	1	5%
Privacy and ethical considerations	1	5%
AI capabilities and insights	4	20%
User experience and interface	6	30%

Fig. 15 The managers shared positives, negatives, and potential improvements for EmoReflex.

👎 *Negative feedback: User interaction and experience.* The participants shared negative comments on clutter, UI issues, and unnecessary features, and dependence on team's accuracy and honesty.

[MP8] said that "too much clutter - bad UI - unnecessary features", while [MP15] said "it might need some training or guide to get started and might take some time to get comfortable with it other than that it looks pretty nice!". [MP16] pointed out that "emotions are subjective, and interpreting emotional data may vary from person to person. Some team members may find it challenging to express emotions accurately" and the insights "depend on the team's accuracy and honesty" [MP1].

👎 *Negative feedback: Other concerns and resistance.* Some participants shared other concerns and resistance. These included concerns about micromanaging and privacy, individual perspective and skepticism, loss of concentration and boredom, potential managerial resistance, resistance to change and user adaptation, and a few more drawbacks and considerations as detailed below.

[MP5] said "not sure if they would be personally fine with this level of personal micromanaging, though"; [MP14] said "I could see some workers being against this level of micromanaging and somewhat privacy invasive", and [MP18] said "it may risk privacy concerns and over-reliance on technology for interpreting complex human emotions and dynamics". [MP4]'s individual perspective and skepticism was that it is necessary to "communicate with your team directly instead of trying to supplement human interaction with this type of "emotions" aggregation. I'm begging you, use your words!". [MP19] raised the concern of loss of concentration and boredom as "but otherwise it can easly get bored and loss your concentracion". However, they did not comment further on this. [MP11] noted the potential managerial resistance. They said "some senior managers may not see the value in monitoring the mood of employees and may not want to invest in such tools". [MP20] stated "a possible con is adapting it and that the user wants to implement this option in their usual tasks, there is resistance to the use of new tools". They fruther commented, "The possible con is to achieve a change in people's mentality to want to adapt the use of new tools to use in their activities". This resistance to change and user adaptation, could

be normal to any new tool. In addition to that some other drawbacks and concerns were also noted by the participants such as "it also comes with drawbacks like cost, time commitment, privacy concerns, and being not a cure-all solution. Ultimately, the decision depends on your specific needs and goals, carefully weighing the pros and cons against alternative solutions" [MP3], "EmoReflex collects data on users' facial expressions, heart rate, and other physiological responses. It is important to ensure that this data is collected and used ethically and responsibly" [MP17]. [MP17] further commented " Cost: EmoReflex can be expensive to implement and maintain. It is important to weigh the costs against the potential benefits before making a decision about whether or not to use it. Not a magic bullet: EmoReflex is not a magic bullet for solving all workplace problems. It is important to use it in conjunction with other strategies for improving emotional intelligence and workplace culture".

Potential improvement: Confidence and transparency. [MP16] shared their opinion on confidence and transparency. As suggested by [MP16] "to address privacy concerns, EmoReflex could implement granular privacy settings. Users should have control over which emotional data they share and with whom can be done".

Potential improvement: Privacy and ethical considerations. [MP17] provided their suggestion on improving the privacy and ethical considerations. [MP17] stated improvements in "Transparency and control: Provide users with more transparency about how their data is collected, used, and stored. Offer options for opting out of specific data collection or deleting personal information; and anonymization and aggregation: Explore ways to anonymize or aggregate data while still enabling effective training and insights" can be done. Further they said, " Ethical considerations: Clearly outline the ethical principles guiding data collection and usage, addressing potential biases and discrimination risks".

Potential improvement: AI capabilities and insights. Some participants stated "expansion of AI capabilities to further drive efficiency" [MP2] as a potential improvement. For example, "recognising subtler emotions: Beyond basic moods, understanding frustration, anxiety, or cultural nuances would offer more personalized guidance; culturally aware approaches: tailoring feedback and recommendations to different cultures could boost effectiveness in diverse teams; and actionable advice: Suggesting concrete steps based on individual profiles and situations would empower users to manage emotions in real-time" [MP3]. Also, "Enhanced accuracy in emotional and performance metrics interpretation and stronger privacy safeguards to protect team member data and trust" [MP18].

Potential improvement: User experience and interface. The participants shared how EmoReflex can be improved in terms of user onboarding and experience, and user experience and interface.

As [MP20] said, "It is very helpful if the environment is simple and clear enough to quickly identify the tools that need to be used at the moment; perhaps it would be a good option to improve the aesthetic details to make the environment more user-friendly. Although it is a simple enough environment to identify what needs to be done, it could greatly improve its aesthetic appearance to make it more pleasant for the user". [MP15] and [MP6] gave similar feedback respectively: "May be more interactive webpage and bit more color needs to added so that it will help differentiate different

tasks"; "The UI might be much more informative and intitutive". With respect to user onboarding, [MP1] suggested to "give 1 months free time to a new users to allow them to see the benefit of it". This comment would be useful when it comes to the commercialisation stage of the tool.

5 Discussion

5.1 Implications for Design

Implications derived from quantitative findings. The SUS scores resulted by both user groups indicate the current state of the tool falls under grade B, indicating that there are improvements that can be done to make the tool more usable or to bring it further to grade A. This was clearly complimented by the qualitative findings.

The mental work load test result for the use case of reflection box indicated a moderate frustration level for the developers. In improving the tool, options such as switching on and off the features could be implemented to avoid such frustrations when using the tool.

Implications derived from qualitative findings. Some of the qualitative findings were repetitive among both user groups. Therefore, we further analysed the findings to derive common categories for positives, negatives, and potential improvements. The common findings on positives can be used for marketing purposes and the negatives an improvements can be used for improving the tool in the next cycle of development.

👍 Positives: Team collaboration and dynamics were identified as a positive of EmoReflex by both user groups. While the managers pointed out EmoReflex's ability of enhancing team dynamics, the opinions shared by developers noted the managerial decision support and collaboration that EmoReflex can contribute to. It is interesting to see how the two user groups identified the positives of each other. Both user groups also said user experience – efficient and user centric as shared by developers and user-friendly experience as shared by managers as a positive of EmoReflex. However, we found the same as a negative and and as an improvement of EmoReflex too, concluding the mix opinions around the user experience of EmoReflex.

👎 💡 Negatives and potential improvements: User interaction and experience as a negative was pointed out by both user groups as mentioned earlier. The developers shared that there could be communication and feedback preferences, and managers shared negative comments on user interaction and experience. Both user groups shared that user experience could be improved. They also shared confidence and transparency in the tool also can be improved.

Since some findings were reported only by a limited number of participants, it is important to collect further data to validate the findings to inform the design decisions of the tool.

In any given user study, the feedback are subjective. Also, a tool is never a replacement for human interaction. Therefore, both of these points have to kept in mind during further development of the tool and especially the second point has to be kept in mind when using the tool in the real world.

5.2 Limitations

The user study was conducted with a set of 40 potential users only. Further feedback is required to have concrete findings for design implications. The approach used for the user study was a survey-based approach. In an industry setting, the prototype would be released as a beta version for potential users to use for a certain period to provide us with feedback. The feedback may differ if the participants used the tool themselves rather than grasping the use cases through videos and answering questions through a form. in conjunction with the real use of the tool, methods such as interviews and focus groups also could be conducted in the future to derive solid findings.

EmoReflex is currently implemented using infrastructure that are available for free to use as it is for prototype purpose only. We did not investigate the compliance of the infrastructure because of that. At the point of commercialisation, the tool is likely to be use subscriptions of the same or migrated to different infrastructure after careful examination of the compliance certifications they have such as SOC2.

In Emotimonitor (the previous version), we used the well-established discrete emotion schema by [7]. However, given the feedback we received for Emotimonitor, in the current version of EmoReflex, we decided to have the option for the teams to define the emotions they would like to have in the tool. A default set of emotions that is backed by a validated questionnaire/emotion schema could be provided in the tool. In addition to that, a set of other recommended emotional schemas could also be provided for the teams to consider, and even the self-assessment manikin [8]. The use of emojis to represent emotions was validated in Emotimonitor as the feedback we received from potential users was positive. Hence, we have used emojis in EmoReflex too. However, this could be changed/left the same in future iterations based on the feedback the potential users are going to provide.

For further improvement of EmoReflex, more research and management science specialists will be consulted to strengthen experimental results and obtain more precise recommendations for the design.

More managers provided positive feedback for EmoReflex than the developers. This could be that, managers may appreciate the tool because it helps them monitor the team's emotional state and improve overall well-being. But developers could be concerned about being monitored, and it might be reassuring to show them positive feedback from managers, who are primarily interested in promoting team well-being rather than micromanaging individual emotions. Therefore, from the practical use point of perspective, we do not encourage any software team to entirely rely on EmoReflex but use it only as a supportive tool to function better. The managers and the teams should maintain the human interactions they have while using EmoRlex upon the agreement of both parties.

To use a tool like EmoReflex at a workplace where sensitive information such as emotions is captured and shared, the consent of all team members is necessary. This is vital when it comes to negative emotions. However, in an emotionally intelligent environment, where developers are keen on sharing how they feel with their manager when working on tasks [9], exposing who felt what would not be seen as an issue, but rather as a sign of high emotional awareness. However, this might be not the case for every software team. Hence, EmoReflex anonymises the user information so that

the manager does not know which developer felt which emotion. But, in a small team where one developer is assigned to one task, then the identity of the developer could be revealed to the manager. In such cases, the developer is free to opt-out sharing what they feel through EmoReflex as the features of EmoReflex are not mandatory to use, but use as preferred.

While ChatGPT enhances the recommendation function in EmoReflex by providing insights, its limitations should be clearly acknowledged. AI models like ChatGPT may introduce biases based on their training data, potentially misinterpreting nuanced emotions or cultural differences. Additionally, the model's responses are influenced by prompt phrasing and lack true emotional understanding, which can affect the accuracy of its analysis. It may also struggle with context-specific subtleties, leading to generalised suggestions. These limitations highlight the importance of human oversight to validate the recommendations and ensure they align with the intended emotional context and team dynamics.

The scales (NASA Task Load Index and SUS) we used in the user evaluation are validated, well-established scales, and widely used. The rest of the questions were standard user study questions one would ask in industry contexts. However, we recruited participants batchwise and also we had a question for the participants to provide any other feedback they have. We checked the data upon every recruitment to understand if our questionnaire was valid. We did not come across any issues.

6 Related Work

The tool built for emotional awareness in software contexts is extremely limited. This includes our very own work which was a pioneer in developing such tools.

EmoReflex is an improved version of our previous tool, Emotimonitor [6]. Emotimonitor only had a emotion reporting component where as EmoReflex allows users to report their emotions against different metrics and enter reflections using the Trello plugin. Emotimonitor did not have a configuration function for the managers to setup emotions or metrics whereas in EmoReflex, the managers can define them. Emotimonitor also had a dashboard which was given as a part of the Trello plugin allowing managers to see the emotions across tasks and time in a given board. EmoRflex has a dashboard hosted separately allowing managers to select the projects and see further insights including the status of the tasks and also the emotions of the developers along with AI generated recommendations to guide them handle the team better.

Grassi et al. [10] built a proof of concept of a tool which visualise the self-reported emotions, activities, and their biometrics during agile retrospective meetings. In comparison to this, EmoReflex rather focuses on emotions in the development process but not during agile ceremonies/ meetings. We also did not integrate capturing and reporting of developer biometrics into our tool. In future, if such features are to be implemented, user consent has to be taken before using the tool in the real world. part from the above two, to the best of our knowledge, there exists a few work for software engineering contexts that used emotion extraction from text such as natural language messages and comments [11–13].

Some studies in software engineering have explored emotion extraction from text, providing valuable insights into the emotional dynamics within teams. The development of EmoTxt demonstrates the use of machine learning to analyze emotions in software engineering communications, highlighting the potential to inform better team management through automated emotional insights [13]. Additionally, methods for sentiment strength detection in short texts such as commit messages and code reviews offer a nuanced understanding of sentiment in digital communications that can enhance team well-being and productivity [5]. Moreover, the study on extracting emotions from GitHub interactions suggests that recognizing emotional contexts in development activities can improve collaborative aspects and project outcomes [14]. These approaches underscore the importance of integrating emotion-aware systems in software development to enhance team interactions and project management.

Similar to EmoReflex, studies have also been conducted on automating data collection and analysis to support user engagement and skill development. ConstelaDev [15] combines special techniques, including natural language processing (NLP), to understand students' feelings and progress in a bootcamp setting. This approach helps maintain a safe and collaborative environment. While EmoReflex focuses on software and developer teams, ConstelaDev is designed to be used in various educational settings.

Recent advancements in emotion recognition technology have significantly enhanced requirements engineering (RE) by integrating emotional awareness into software development. Studies have demonstrated the potential of using biometric sensors and facial expression analysis for recognizing and understanding emotions during RE activities. For example, Novielli [16] used wearable biometric sensors to track emotions during programming tasks, improving task performance and communication. Similarly, Cheng [17] introduced the multi-modal emotion recognition platform (MEmoRE), which captures and analyzes stakeholders' emotional cues through facial expressions, vocal intonation, and textual sentiment analysis in real-time, enhancing the elicitation and validation of requirements.

Teslia and Klevanna [14] developed an intelligent project planning system that automates the creation of Project Schedule Network Diagrams (PSNDs) using a modified reflex method. This system reduces the need for human input by integrating various data sources, such as expert knowledge and historical project data, to produce accurate and realistic project plans. Their approach is similar to how the MEmoRE platform uses multi-modal data to improve decision-making in requirements engineering, highlighting the move towards AI-driven solutions in project management.

7 Summary

Project/task management tools should not only be limited to serve its sole purpose of project/task tracking, but also to support enhance the human aspects of the teams. We developed EmoReflex as a supportive tool for software teams to improve their emotional awareness and function better as a team. This paper reports the insides of the tool at its early stages followed by a user study conducted with forty potential users.

Acknowledgements. Madampe and Grundy are supported by ARC Laureate Fellowship FL190100035. Our sincere gratitude goes to the participants who took part in the user study.

Author Contribution

Kashumi Madampe: Conceptualization, Supervision, Project Administration, User Study Conduction, Writing - Original Draft, Review & Editing. John Grundy: Conceptualization, Supervision, Writing - Review & Editing. Minh Nguyen: Writing - Original Draft, Software Development. Ellen Welstead-Cloud: Project Administration, Software Development. Vinh Tuan Huynh, Linh Doan, William Lay, Sayed Hashim: Software Development.

References

[1] Madampe, K., Hoda, R., Grundy, J.: The emotional roller coaster of responding to requirements changes in software engineering (2021)

[2] Colomo-Palacios, R., Hernández-López, A., García-Crespo, Soto-Acosta, P.: A study of emotions in requirements engineering. (2010). https://doi.org/10.1007/978-3-642-16324-1_1

[3] Graziotin, D., Wang, X., Abrahamsson, P.: Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering. Journal of Software: Evolution and Process (2015) https://doi.org/10.1002/smr.1673

[4] Kolakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wrobel, M.R.: Emotion recognition and its application in software engineering. (2013). https://doi.org/10.1109/HSI.2013.6577877

[5] Müller, S.C., Fritz, T.: Stuck and frustrated or in flow and happy: Sensing developers' emotions and progress. (2015). https://doi.org/10.1109/ICSE.2015.334

[6] El-Migid, M.-A.A., Cai, D., Niven, T., Vo, J., Madampe, K., Grundy, J., Hoda, R.: Emotimonitor: A trello power-up to capture and monitor emotions of agile teams. Journal of Systems and Software **186**, 111206 (2022) https://doi.org/10.1016/J.JSS.2021.111206

[7] Harmon-Jones, C., Bastian, B., Harmon-Jones, E.: The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. PLoS ONE **11** (2016) https://doi.org/10.1371/journal.pone.0159915

[8] Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry **25**(1), 49–59 (1994)

[9] Madampe, K., Hoda, R., Grundy, J.: Supporting emotional intelligence, productivity and team goals while handling software requirements changes. ACM Transactions on Software Engineering and Methodology **33**(6), 1–38 (2024)

[10] Grassi, D., Lanubile, F., Novielli, N., Serebrenik, A.: Towards supporting emotion awareness in retrospective meetings. In: 2023 IEEE/ACM 45th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), pp. 101–105 (2023). https://doi.org/10.1109/ICSE-NIER58687.2023.00024

[11] Werder, K., Brinkkemper, S.: MEME - Toward a Method for EMotions Extraction from GitHub. In: 2018 IEEE/ACM 3rd International Workshop on Emotion Awareness in Software Engineering (SEmotion), pp. 20–24 (2018)

[12] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology **61**(12), 2544–2558 (2010) https://doi.org/10.1002/asi.21416 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21416. Accessed 2020-08-31

[13] Calefato, F., Lanubile, F., Novielli, N.: EmoTxt: A toolkit for emotion recognition from text. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops And Demos (ACIIW), pp. 79–80 (2017). https://doi.org/10.1109/ACIIW.2017.8272591

[14] Teslia, I., Klevanna, G.: Development of a reflective intelligent project planning system. In: ITPM, pp. 170–182 (2022)

[15] Perazzoli, S., Santos, W., Santana Neto, J.P.: Towards a systemic-based automated platform designed to foster collaborative and co-responsibility environment. Available at SSRN 4249435 (2022)

[16] Novielli, N., Grassi, D., Lanubile, F., Serebrenik, A.: Sensor-based emotion recognition in software development: Facial expressions as gold standard. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8 (2022). https://doi.org/10.1109/ACII55700.2022.9953808

[17] Cheng, B., Arora, C., Liu, X., Hoang, T., Wang, Y., Grundy, J.: Multi-modal emotion recognition for enhanced requirements engineering: A novel approach. In: 2023 IEEE 31st International Requirements Engineering Conference (RE), pp. 299–304 (2023). https://doi.org/10.1109/RE57278.2023.00039

Appendix A Developer Feedback - qualitative - detailed

	High level theme	Theme	Participants	
Positives	Customisation and metrics	Customisable metrics	1	5%
	Supports professional development	Professional development and training insights	1	5%
	Effective task management	Enhanced task understanding	2	10%
		Data utilisation	1	5%
		Improved task assignment	1	5%
		Strategic rating and retrospective facilitation	1	5%
	Managerial decision support and collaboration	Informed managerial decision-making	1	5%
		Managerial decision support	1	5%
		Project manager agrement	1	5%
	Efficient and user-centric	Efficiency and user-friendly	2	10%
		Adaptable for remote work dynamics	1	5%
		Enhanced team communication	1	5%
		Playful and easy-going	1	5%
		Transparent communication	1	5%
Negatives	Communication and feedback preferences	Desire for more information and clarity	2	10%
		Anonymity and openness	1	5%
		Preference for traditional feedback	1	5%
	Complexity and tool dependency	Concerns about complexity and dependency on tools	2	10%
	Concerns and resistance	Annoyance and lack of consistent usage	1	5%
		Concerns about time and engagement	1	5%
		Privacy concerns	1	5%
		Resistance to mandatory usage	1	5%
		Skepticism about emotional aspect	1	5%
		Uncertainity about benefits	1	5%
Potential improvements	Confidence and transparency	Confidence and transparency	1	5%
	Flexibility and optional features	Optional features and flexibility	2	10%
	Functionality and metrics	Functionality and metrics	3	15%
	Satisfaction and purpose	Overall satisfaction and functionality	4	20%
		Pragmatic acceptance	1	5%
		Uncertainity and resistance	1	5%
	User experience and interface	User interface and ease of use	4	20%
		Aesthetic and emotional component	3	15%

Fig. A1 The developers shared positives, negatives, and potential improvements for EmoReflex.

Appendix B Manager Feedback - qualitative - detailed

	High level theme	Theme	Participants	
Positives	Emotional intelligence and well-being	Emotion tracking and AI utilisation	2	10%
		Enhanced emotional intelligence and well-being	2	10%
		Innovative and fun approach	1	5%
	Enhanced team dynamics	Team insights and at-a-glance view	4	20%
		Task management and organisation	3	15%
		Real-time insights for team management	1	5%
	User-friendly experience	User-friendly and intuitive	5	25%
		Positive feedback on simplicity	1	5%
Negatives	User interaction and experience	Clutter, UI issues, and unnecessary features	3	15%
		Dependendence on team's accuracy and honesty	2	10%
	Other concerns and drawbacks	Concerns about micromanaging and privacy	3	15%
		Drawbacks and considerations	2	10%
		Individual perspective and skepticism	1	5%
		Loss of concentration and boredom	1	5%
		Potential managerial resistance	1	5%
		Resistance to change and user adaptation	1	5%
Potential improvements	AI capabilities and insights	AI capabilities and insights	4	20%
	Confidence and transparency	Confidence and transparency	1	5%
	Privacy and ethical considerations	Privacy and ethical considerations	1	5%
	User experience and interface	User onboarding and experience	5	25%
		User experience and interface	1	5%

Fig. B2 The managers shared positives, negatives, and potential improvements for EmoReflex.