UNIVERSITY OF TARTU
Institute of Computer Science
Data Science Curriculum

Marti Taru

Goal-Setting in Primary School with LLM Personalised Feedback: A Pilot Study

Master's Thesis (15 ECTS)

Supervisor:
Kristjan-Julius Laak, MSc

Tartu 2025

Goal-Setting in Primary School with LLM Personalised Feedback: A Pilot Study

Abstract

Background: Goal-setting is a key element of self-regulated learning (SRL) and metacognitive development in primary education. Large Language Models (LLMs) can provide timely, personalized feedback that may enhance these skills, but evaluating their impact in classroom-like settings requires practical and valid measurement approaches.

Objective: This pilot study examined whether readily available indicators—such as text length, self-assessment scores, and input similarity—can detect patterns consistent with enhanced reflection and SRL when pupils receive LLM-based feedback on cross-curricular goals (e.g., behaviour, collaboration, questioning) intended to support more effective study processes and improved learning outcomes.

Methods: Thirty-one pupils aged 10–13 participated in a six-week repeated cross-sectional study using a custom web application. Pupils were randomly assigned to a test group (receiving GPT-4o feedback) or a control group (no feedback). Weekly goal-setting and progress-assessment inputs were analysed using independent-samples t-tests, Kolmogorov–Smirnov tests, and descriptive statistics. Post-experiment interviews with pupils and teachers provided qualitative context.

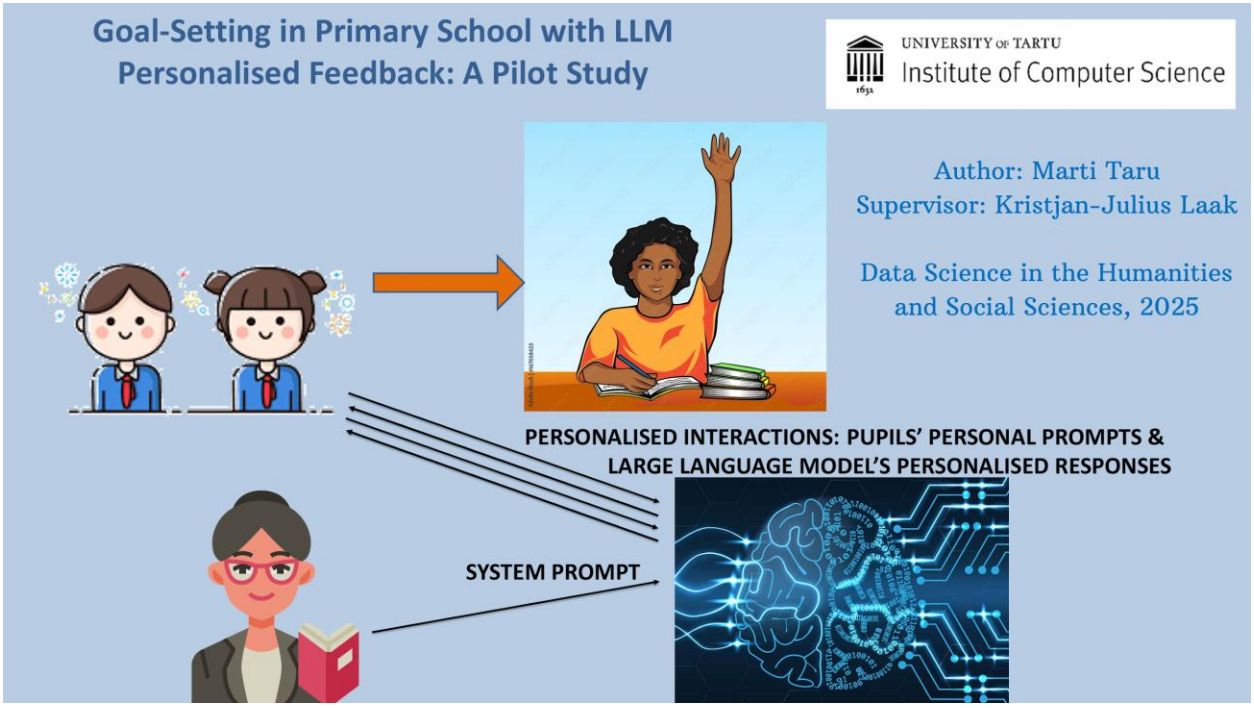
Results: The test group generally produced longer goal statements and progress assessments and reported slightly higher self-assessed weekly progress marks. One session showed a statistically significant difference in progress-assessment length favouring the test group; other sessions showed similar but non-significant trends. No significant differences were found in lexical or semantic input similarity. Qualitative findings indicated that both pupils and teachers perceived the feedback as motivating and useful.

Conclusions: Readily available indicators revealed patterns suggestive of more elaborate reflection with LLM feedback, but statistical significance was rare—likely due to small sample size, control-group attrition, and limits of the chosen measures. The findings support the feasibility of using simple metrics as initial indicators, while underscoring the value of longitudinal designs, refined reflection-quality measures, and larger, more representative samples in future studies.

Keywords: self-regulated learning, metacognition, goal-setting, large language models, educational technology, primary education, pilot study

CERCS: S281 Computer-assisted education

Visual Abstract



Eesmärgistamine põhikoolis personaliseeritud LLM-i tagasiside abil: pilootuuring

Kokkuvõte

Taust: Eesmärkide seadmine on algkoolis enesereguleeritud õppimise (SRL) ja metakognitiivse arengu keskne komponent. Suured keelemudelid (LLM-id) pakuvad uusi võimalusi vahetu ja isikupärastatud tagasiside andmiseks, mis võiks neid oskusi toetada, kuid nende mõju hindamine koolikeskkonnas nõuab praktilisi ja usaldusväärseid mõõdikuid.

Eesmärk: See pilootuuring uuris, kas lihtsad ja kergesti kogutavad indikaatorid – näiteks teksti pikkus, enesehindamise tulemused ja sisendi sarnasus – suudavad tuvastada mustreid, mis viitavad paremale refleksioonile ja SRL-ile olukorras, kus õpilased saavad LLM-põhist tagasisidet oma õppe- ja käitumiseesmärkide kohta.

Meetodid: 31 õpilast vanuses 10–13 osales kuue nädala pikkuses korduvate läbilõigete uuringus, kasutades spetsiaalselt loodud veebirakendust. Õpilased määrati juhuslikult testgruppi (LLM-tagasiside GPT-4o kaudu) või kontrollgruppi (ilma tagasisideta). Iganädalased eesmärkide seadmise ja edusammude hindamise sisendid analüüsiti sõltumatute valimite t-testide, Kolmogorovi–Smirnovi testide ja kirjeldava statistika abil. Eksperimendijärgsed intervjuud õpilaste ja õpetajatega pakkusid kvalitatiivset lisakonteksti.

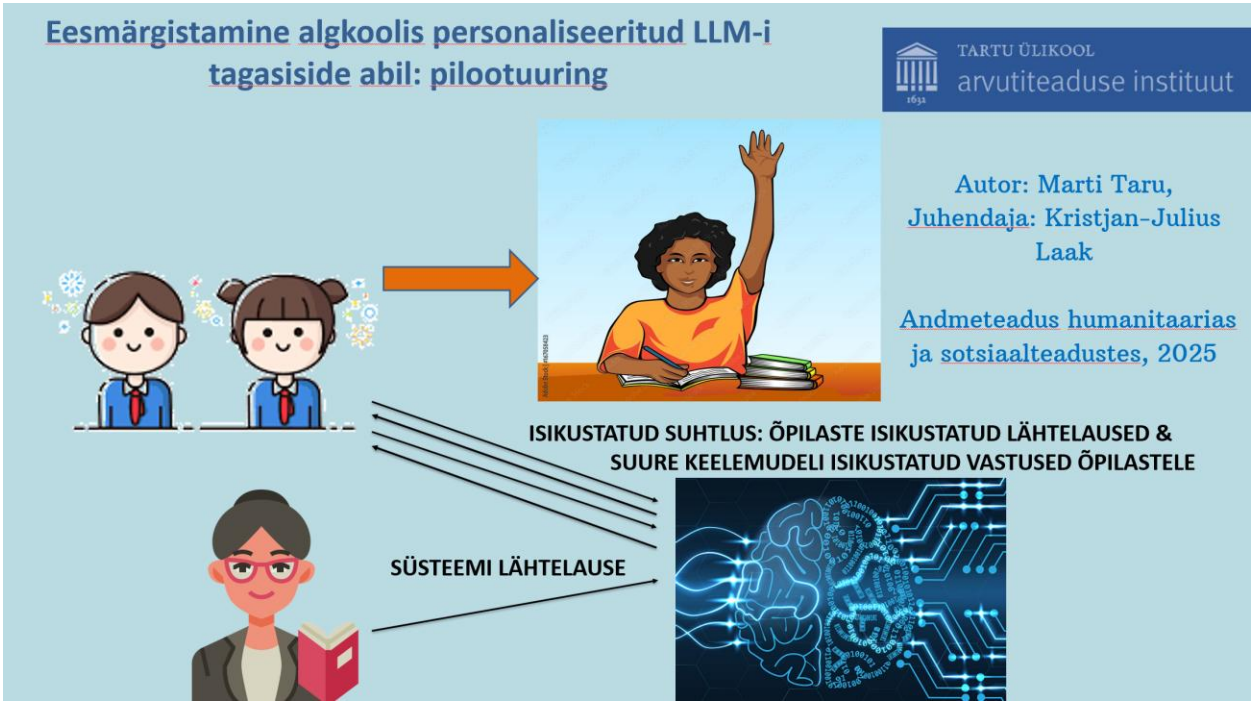
Tulemused: Testgrupp esitas üldiselt pikemaid eesmärgipüstitusi ja edusammude hindamisi ning andis veidi kõrgemaid enesehinnanguid iganädalastele edusammudele. Ühe sessiooni puhul oli edusammude hinnangute pikkuse erinevus statistiliselt oluline testgrupi kasuks; teistes sessioonides esines sarnane, kuid statistiliselt mitteoluline muster. Leksikaalse või semantilise sarnasuse osas olulisi erinevusi ei leitud. Kvalitatiivsed tulemused näitasid, et nii õpilased kui ka õpetajad pidasid LLM-tagasisidet motiveerivaks ja kasulikuks.

Järeldused: Lihtsad indikaatorid näitasid mustreid, mis viitavad põhjalikumale refleksioonile LLM-tagasiside toel, kuid statistiline olulisus jäi harvaks, tõenäoliselt väikese valimi, kontrollgrupi suure väljalangevuse ja mõõdikute piirangute tõttu. Uuring toetab lihtsate mõõdikute kasutamist esmaste signaalidena ning rõhutab vajadust kasutada longituudset uuringudisaini, täpsemaid refleksiooni kvaliteedi mõõdikuid ning suuremaid ja esinduslikumaid valimeid edasistes uuringutes ja rakendustes.

Märksõnad: enesereguleeritud õppimine, metakognitsioon, eesmärkide seadmine, suured keelemudelid, haridustehnoloogia, põhiharidus, pilootuuring

CERCS: S281 Arvuti õpiprogrammide kasutamise metoodika ja pedagoogika

Visuaalne kokkuvõte



Acknowledgements

I express my sincere gratitude to my supervisor, Kristjan-Julius Laak, for his encouragement to undertake this project and for his support throughout its completion, as well as for acting as a vital connection between me and the school. His contribution was threefold: providing feedback into the conceptual section; guiding use-case development and frontend design and development; and establishing contacts with teachers and pupils while successfully promoting participation in the experiment.

I am especially indebted to Sergey Kovalskiy, an experienced full-stack developer who provided targeted mentoring and technical support at critical points in the project. His role was to help me overcome specific technical challenges, review my work, and suggest approaches while I retained responsibility for implementation. His input was invaluable in guiding development and helping me progress past technical bottlenecks.

I also acknowledge the support of advanced AI language models, which were used to review code and text and to refine the grammar, clarity, tone, and flow of the thesis manuscript.

Contents

1 Introduction

Research Problem

Fostering metacognitive awareness and self-regulated learning (SRL) is a key educational objective in upper primary education, particularly as students begin to take greater responsibility for their learning. Goal-setting plays a central role in SRL by enabling learners to plan, monitor, and evaluate their progress. However, supporting high-quality goal-setting and reflection in classroom settings remains both a practical and pedagogical challenge, primarily because teachers lack the time to give each student timely, personalized feedback on their goals and reflections.

Recent advances in large language models (LLMs) offer new possibilities for delivering automated, individualized feedback that could enhance students' reflective thinking and self-regulatory behavior, helping to address this challenge. Evaluating the impact of such interventions in real-world educational settings is complex. Existing assessment tools in educational research are often ill-suited for capturing reflection quality, metacognitive engagement, or the iterative nature of goal formulation in school settings because they are often too complex and resource-intensive.

The aim of this study is to explore whether simple, scalable metrics can detect the effects of large language model–based feedback on primary school pupils' goal-setting practices. Such an approach could be integrated into the daily operations of schools, rather than being limited to one-off academic studies. By providing teachers with easily interpretable indicators, it has the potential to support instructional decision-making and enhance pupils' learning outcomes.

The study was carried out over six weeks with pupils aged 10–13, using a web application to collect and analyze weekly goal-setting and self-assessment inputs. Pupils in the test group received automated, personalized feedback from GPT-4o; the control group did not. The approach combined quantitative and qualitative methods to examine patterns of engagement and reflection.

The research questions are presented after the theoretical and conceptual framework to ensure they are grounded in clearly defined concepts and contexts, allowing for precise formulation and alignment with established literature.

2 Theoretical and Conceptual Framework

2.1 Metacognition and Self-Regulated Learning

In educational research, metacognition and self-regulated learning (SRL) are considered significant concepts because they are key drivers of effective learning and long-term academic success. Both are linked to improved achievement, deeper understanding, and better transfer of knowledge across contexts. They are also vital for lifelong learning, as they equip students to independently manage their learning beyond formal instruction. These constructs serve as strong predictors of academic outcomes and are central to designing interventions that enhance student autonomy and learning efficiency. On a deeper level, they help explain how students learn effectively and why some achieve better results even under similar instructional conditions (Zimmermann, 2004; Kersna et al., 2025). As Georghiades (2004) notes, these concepts provide a broad foundation for measuring reflection quality and goal-setting effectiveness. A related blend of metacognition and learning is captured in the notions of metalearning, deutero-learning, and mindfulness, which refer to an awareness of problems, situations, and the ways in which they are thought about and discussed in educational contexts.

Metacognition refers to individuals' awareness and regulation of their own cognitive processes. It encompasses two core components: metacognitive knowledge—awareness of one's cognitive abilities, strategies, and the conditions under which they are effective—and metacognitive regulation—the ability to plan, monitor, and evaluate one's learning or problem-solving activities. Metacognition is central to effective learning, shaping how learners set goals, choose strategies, and adapt to challenges. It is related to, but distinct from, self-regulated learning (Veenman, Van Hout-Wolters & Afflerbach, 2006).

Self-regulated learning is the process by which learners take active control of their own learning experience. It involves setting goals, selecting appropriate strategies, monitoring progress, and adjusting approaches based on feedback. Self-regulated learners are proactive, managing their motivation, behavior, and learning environment to optimize outcomes (Kurt, 2023).

2.2 Large Language Models in Educational Settings

Educational research has established that formative feedback is crucial for learning effectiveness. According to Shute (2008), formative feedback is defined as information provided to learners during learning with the aim of improving their performance. It is non-evaluative, supportive, and focused on guiding learners toward the correct understanding or skill. Its purpose is to reduce the gap between current performance and desired learning goals, typically through timely, specific, and actionable guidance. However, traditional classroom settings often limit teachers' ability to provide formative feedback at scale. This challenge has driven interest in automated feedback systems that can provide immediate, personalized responses to student work. Recent research demonstrates promising applications of LLMs in supporting the development of metacognitive skills, with even the majority of studies employing strict experimental designs reporting predominantly positive results (Maier & Klotz, 2022).

Within the framework of self-regulated learning, goal formulation is widely recognized as a critical first step. A recent study by Martins Van Jaarsveld et al. (2025) found that LLM-generated feedback had a delayed but cumulative effect on the quality of students' goals across multiple iterations. The combination of guidance and feedback produced the strongest improvements in goal quality, while guidance alone led to initially high-quality goals that declined over time. These findings suggest that personalized feedback may play a key role in sustaining and enhancing the quality of students' learning objectives over time.

Wang (2025) found that fifth-grade students using AI chatbots for visual programming tasks showed significant improvements in programming self-efficacy, particularly in their perceived ability to manage and direct their own learning processes. The personalized support provided by the AI system reduced barriers while encouraging greater ownership of the educational journey. Li et al. (2024) demonstrated that InquiryGPT systems using Predict-Observe-Explain-Evaluate processes enhanced higher-order thinking through dialogic feedback mechanisms in STEM education settings. Students showed improved ability to monitor their own understanding and adjust learning strategies accordingly.

Self-regulated learning capabilities show consistent improvement when students engage with well-designed LLM systems. Lin et al. (2024) found that fifth-grade students using GPT-4 error analysis

in mathematics demonstrated significant gains in problem-solving skills and mathematical confidence, with particularly strong effects for initially low-achieving students. The system's focus on error identification and correction enhanced students' ability to monitor their own mathematical reasoning.

Guo et al. (2025) observed significant improvements in STEM skills when LLM systems were integrated into submarine engineering projects for middle school students. The project-based context enhanced students' scientific knowledge, computational thinking, and problem-solving abilities while developing self-regulatory skills necessary for managing complex, extended learning tasks.

Kazemitabaar et al. (2023) provided crucial insights into optimal conditions for educational LLM deployment through their study of 10-17-year-old students learning Python programming. Their research revealed that hybrid approaches combining AI assistance with active student engagement produced superior outcomes compared to over-reliance on AI-generated solutions. This emphasizes the importance of maintaining student agency in the learning process.

The number and breadth of LLM-supported applications have been expanding at a fast pace. In the broad universe of LLM-supported learning applications, this application and study fall into the category of content personalization in the broader category of adaptive learning applications (Wang et al., 2024).

2.3 Research Question and Approach

Drawing on theoretical frameworks of metacognition, self-regulated learning and metalearning, the study explores whether readily available metrics might show detectable patterns consistent with enhanced metacognitive engagement and self-regulatory behavior in primary school. More specifically, the research focuses on extra-curricular goals in primary school settings, such as keeping a tidy desk, helping classmates, or completing homework on time. Pupils at the study school have practiced such goal setting for years, but without LLM assistance.

LLM-assistance in such goal setting has not been explored earlier, even though IT solutions and also LLMs have been widely used in educational settings over decades. Hence, the study focuses on exploring the role of LLM-feedback on pupils’ practice of setting extracurricular goals in a school setting.

Many studies investigating LLM-supported learning rely on complex, resource-intensive sets of indicators, often requiring detailed multimodal data, sophisticated analysis pipelines, and large research teams (see section 2.2). Such approaches, while powerful, are often impractical for smaller research groups or for routine use in schools. This study therefore focuses on identifying and testing readily available indicators – metrics that can be extracted and analyzed with minimal technical and logistical demands – while still providing meaningful insights into pupils' engagement with LLM-supported feedback.

The primary **research question** asks: **Can readily available indicators provide meaningful insights into how pupils engage with LLM-supported feedback?**

The main research question is split into four **research statements** that explore four specific patterns using readily available metrics.

Research statement 1: The test group uses more characters in goal wording than the control group, potentially indicating deeper metacognitive engagement in goal formulation.

Research statement 2: The test group reports higher weekly progression marks than the control group, suggesting enhanced self-monitoring.

Research statement 3: The test group uses more characters in weekly progress assessments than the control group, potentially reflecting increased metacognitive reflection.

Research statement 4: The test group shows larger changes between subsequent inputs than the control group, suggesting adaptive self-regulatory behavior in response to LLM feedback.

Interdisciplinary Approach

Taken together, the theoretical background and research question reflect the interdisciplinary nature of this study, which combines perspectives and methods from both computer science and educational science. The design and implementation of the web-based goal-setting application, including the integration of LLM-generated feedback, are rooted in computer science and human-computer interaction. In contrast, the conceptual framing of goal-setting behavior and the pedagogical implications of feedback are grounded in educational theory and learning sciences. This dual foundation reflects the practical reality of technology-enhanced learning interventions and underscores the importance of combining technical innovation with educational insight.

Bridging these domains in a measurable way relies on principles from statistics and experimental design, which provide the methodological backbone for evaluating intervention effects.

3 Research Design and Methodology

3.1 Ethical Approval and Consent

Before the study began, school management, teachers, all parents or legal guardians, as well as the pupils themselves, were informed of the nature of the experiment, the data collection procedures, and their right to voluntary participation. Written parental consent and pupil assent were obtained in accordance with ethical standards. The study received formal approval from the University of Tartu Ethics Committee (protocol code: 388/T-4, date of approval: September 2024).

3.2 Study Design

To explore the challenges of measuring the effects of LLM-generated feedback on pupils' goal-setting abilities, the study employed a **multimethod design** that combined structured digital data collection with brief post-experiment interviews. Quantitative data were gathered via a custom-built web application over a six-week period, during which participants were randomly assigned to either a test group (receiving personalized LLM feedback) or a control group (no feedback). Each week, pupils submitted their goal-related reflections through the application.

In addition to the structured digital data, qualitative insights were obtained through short interviews with both pupils and teachers. While exploratory and limited in depth, these interviews provided valuable contextual information that supported interpretation of the quantitative findings and informed recommendations for further development of the intervention and application.

3.3 Participant Recruitment and Group Assignment

The study was conducted in a primary school in Tallinn that actively adopts recent advances in teaching methodologies. All pupils in the school were included in the sample.

Participants were randomly assigned to test and control groups using a random number generator. Specifically, the process was completed using MS Excel function round(RANDBETWEEN(0,1), 0) in an MS Excel spreadsheet table. This was a pragmatic solution as the list of participants' first name was forwarded in the MS Excel format and using randomisation straight on the list was convenient and did not compromise the goal. No stratification was used as there was no additional background information available.

3.3.1 Group Conditions

The application's frontend presented both groups with user interfaces that by and large had similar visual design and logic. The difference was that the test group pupils were displayed personalized responses generated by GPT-4o after each goal-setting and progress assessment input while the control group pupils completed identical activities but received no feedback. Additionally, there were slight differences in button captions between the groups:

- In the test group, the button labeled "Küsi tagasisidet" ("Ask for feedback") corresponded to "Salvesta" ("Save") in the control group.
- The test group's button "Täienda analüüsi tagasiside põhjal" ("Improve your analysis based on the feedback") was matched with "Muuda" ("Change") in the control group.

3.3.2 Personalized Feedback

Pupils' inputs were forwarded to the OpenAI model GPT-4o, which returned personalized responses based on the prompt. Prompts consisted of two parts: a system prompt and a user prompt. The system prompt described the role of a teacher and assigned this role to the LLM; it also incorporated the pupil's personal goal. This remained constant for all users (see Appendix 1). User prompts were the concrete inputs provided by the pupils, varying by input and user.

The wording of the system prompt, described above, was developed through iterative trial and error during preparatory work. Several approaches were explored, including SMART goals, problem-based learning feedback, and other instructional frameworks. Ultimately, a formative feedback-based system prompt was selected.

3.3.3 Session Structure

Each session followed a standardized procedure that began with a pre-treatment phase where users logged into the web application and entered their initial input, which could be either a goal or progress assessment. The very first session had a unique structure: users began by formulating a short goal name with only one entry opportunity and no chance to revise it (this was identical for both groups). This was immediately followed on the same screen by formulation of a longer, more detailed goal, which already incorporated the group differentiation described below. During assessment sessions, participants also provided ratings using a Likert scale. The treatment phase then differentiated between groups: the test group received LLM-generated feedback while the

control group received no feedback at all. In the post-treatment phase, both groups had the opportunity to enter revised or additional input, allowing the cycle to repeat as needed within each session. However, only the test group had feedback to guide their revisions, while the control group made any changes based solely on their own reflection. Users maintained control over session duration and concluded their sessions by clicking a button when they felt they had accomplished enough for that particular session. The complete user experience is documented in Appendix 2, which provides screenshots of all stages users encountered during their first, second, and subsequent logins.

Pupils were not given any training or specific instructions on how to use the web application. Its usage was thought to be self-explanatory and pupils were believed to be interested in and motivated to use the app intensely.

Sessions were not supervised, and the duration of each session was determined by the user. Session duration was not measured and stored in the database.

3.3.4 User Input Analysis Metrics

The UI/UX was designed to support pupils' reflection about their goals and progression toward their individual goals, i.e. they were encouraged to enter a "dialogue" with the LLM (test group) or enter several inputs (control group).

Text length was measured in characters as one readily available metric that might potentially correlate with depth of reflection. This relationship was exploratory though and not validated for this age group or context.

In assessment sessions, pupils also provided self-assessments of weekly progress using a three-category Likert scale.

Input similarity analysis examined semantic and lexical similarity between consecutive inputs to explore patterns in response development.

3.4 Statistical Analysis Plan

3.4.1 Baseline Comparisons

Baseline comparisons were deemed necessary to examine group characteristics at each session. This was important given the variable attendance patterns documented on Figure 2. Pre-treatment input lengths were compared between groups using means and independent samples t-tests to examine any significant differences. Where multiple comparisons were involved, Bonferroni correction was applied. Since a total of 5 assessment sessions took place and the type 1 error per all five sessions was set to 0.05, using Bonferroni correction lowered session-based α to 0.01 (0.05/5).

3.4.2 Exploring Research Statements

For exploring and comparing differences in input lengths as well as in marks (research statements 1, 2 and 3), means were used as metrics summarizing the general tendency of a distribution. In the case of research statements 1 and 3, the count of characters per input served as the input for calculating means. In the case of research statement 2, raw data from participants' progress buttons were used.

For examining the significance of differences in means across research statements 1, 2 and 3, a cross-sectional design was chosen. Alternatively, a panel design could be used and would be preferred from substantive as well as from statistical perspective. However, this approach would result in approximately 50% attrition in the already small control group, creating also severe group imbalance given that 85% of test group participants attended all sessions.

This design approach required the use of independent samples t-tests. The type 1 error per all five sessions was set to 0.05, using Bonferroni correction lowered session-based α to 0.01 (0.05/5).

For obtaining an estimate of similarities of inputs, two different metrics were used:

- Damerau-Levenshtein Distance measures lexical similarity at character level, capturing surface-level changes in text (corrections of typos, punctuation and the like) but unable to distinguish differences in meaning.

- Cosine Similarity measures semantic similarity using embeddings from a multilingual transformer, capturing meaning-level changes while potentially ignoring minor spelling, punctuation and/or formatting differences.

For examining the distribution patterns of similarity scores in research question 4, the two-sample Kolmogorov-Smirnov test was deemed appropriate. This test helps to understand if both observed distributions come from the same underlying distribution, or not.

The overall number of input pairs was limited. The per-session numbers were even lower, making session-by-session analysis impractical. Therefore, this analysis was conducted using a pooled dataset of input pairs, not on a session-by-session basis.

3.5 Post-Hoc Interviews

Following the experiment period when pupils used the webapp, two group interviews were conducted, one with teachers and one with pupils.[1] The interviews lasted between 20 and 25 minutes.

[1] Interviews were conducted in the premises of the school. Neither teachers nor pupils were offered any reward for participation. The interviews were conducted by Kristjan-Julius Laak and recorded. The interviews were transcribed using a VTT application (Tekstiks, https://tekstiks.ee/et), edited and analysed by Marti Taru.

4 System Architecture and Implementation

The data collection instrument was implemented as a full-stack web application based on a standard client-server architecture. More concretely, a 3-layer application architecture was developed, consisting of a presentation layer, an application layer, and a database layer (GeeksforGeeks, 2025; Parker, 2023). The entire system was deployed on a containerized infrastructure.

4.1 System Architecture and the Technology Stack

The technologies for these application layers and the supporting infrastructure were chosen to align with the author's existing expertise in Python and data science, while prioritizing stability, strong community support, and suitability for the project's specific requirements. A summary of the technology stack is provided in Table 1.

The presentation layer

The frontend was built using standard web technologies (HTML/CSS/JavaScript/ReactJS) to ensure the planned user interface features and to create the necessary dynamic and interactive user experience.

The application layer

Python was the natural choice for the backend language due to its prevalence in data science and the author's proficiency.

The Flask framework was selected over more heavyweight alternatives like Django. As a microframework, Flask provides core functionalities like routing and request handling and is relatively simple in development, deployment, and maintenance.

The database layer

MySQL was chosen as the relational database management system (RDBMS). As a powerful, open-source, and highly stable RDBMS, it offers robust data integrity features and a structured query language (SQL) that is ideal for storing the well-defined, tabular data generated by the experiment.

This configuration is well-suited for a single-purpose academic research project like this one. The

frontend component runs in the participant's web browser, rendering the user interface and sending experimental data to the backend. The backend component, running on the server, processes all logic, handles data validation, and manages persistence to the database. All technologies are summarized in the table below.

Table 1. Technology Stack

Component	Technology	Version	Justification for Selection
Operating System	Ubuntu	22.04 LTS	Stable, well-supported operating system, standard for web server deployment.
Containerization	Docker Engine	28.0.4	Enables isolated, reproducible, and portable environments for each service.
Orchestration	Docker Compose V2	2.34.0	Simplifies multi-container development; integrated into Docker CLI.
Language Runtime	Python	3.13.3	Modern, versatile language with extensive libraries and community support.
Backend Framework	Flask	3.0.3	Lightweight Python microframework; aligns with project simplicity and development phase.
Frontend Framework	ReactJS	18.3.1	Modern UI library with component-based architecture, reusable logic, and strong ecosystem.
Frontend	HTML5, CSS3, JavaScript (ES6)	N/A	Standard web technologies providing maximum control over the UI without framework overhead.
Database	MySQL	8.0.41	Open-source relational database suitable for structured experimental data.
Reverse Proxy	Nginx	1.27.5	High-performance server used in the role of reverse proxy for handling HTTPS requests

4.2 UI/UX Design for Pupils

The design of the user interface (UI) and user experience (UX) was not only an aesthetic consideration but a core methodological control. A poorly designed interface could act as a significant confounding variable, leading to participant frustration, boredom, or misunderstanding, thereby compromising the quality and validity of the collected data. Therefore, the application's frontend was designed based on established principles of UI/UX for children, with the explicit goal of maximizing engagement and ensuring the collection of reliable data. The core design philosophy was to create an environment that was intuitive, engaging, and minimized the cognitive load on the 10-13 year old participants (Ux&You, 2023; Aparna, 2024). The complete user experience is documented in Appendix 2, which provides screenshots of all screens users encountered.

4.3 Backend Implementation

The backend of the application served as the central system, responsible for executing the experimental logic, processing participant interactions, and ensuring the secure and reliable persistence of data.

The business server logic was developed using the Flask microframework for Python. Its minimalist philosophy allowed for a simple and transparent project structure, free from the boilerplate of larger frameworks.

The codebase was organized into modules responsible for different concerns. The codebase was organized into modules responsible for different concerns, which included:

- managing the business logic for the experiment and defining API routes,
- making LLM requests,
- handling database interactions,
- ensuring goal formation logic and
- ensuring goal assessment logic.

4.3.1 Custom Concurrency Management

The Flask development server (Werkzeug) lacks built-in support for efficiently handling multiple simultaneous client requests. Although the overall workload for the application was expected to be very low, its interactive nature introduced a high likelihood of concurrent user requests. To

ensure responsiveness under such conditions, a custom concurrency management solution was implemented.

This solution utilizes Python's queue and threading modules (Ramalho, 2022). When a user sends a message, the server responds immediately by placing the message and task metadata into an in-memory queue, keeping the initial HTTPS request lightweight.

A dedicated background worker thread continuously monitors the task queue. It processes tasks sequentially – for example, by invoking goal formation or assessment logic – independent of the main request-response cycle. The queue is limited to a maximum of 20 pending tasks, each corresponding to a user request waiting to be processed. Since the application only queues tasks initiated by user messages, this effectively limits the number of concurrent users whose requests can be processed or are waiting in line. If this limit is reached, additional incoming requests are rejected.

4.3.2 Requests to LLM

Requests to the LLM form the core of the application's functionality, as this is the mechanism through which personalized feedback is generated. Each request sent to the LLM included a system message (defining the LLM's role or the overall context for the interaction), a user message (the participant's specific input), and the goal formulated by the user. The LLM processed the combined information from all of these and returned a response. This LLM-generated response is conceptualized as personalized feedback tailored to each user input. Within the framework of the experimental design, this personalized feedback represents the treatment received by the test group.

The system message instructed the LLM to be an attentive and encouraging teacher who gives feedback to 10-13 year old pupils. It also included the pupil's own goal. Detailed description of the system message is in Appendix 1.

4.3.3 Communication Between the Frontend And Backend: REST API

The communication between the frontend and backend was managed through a Representational State Transfer (REST) Application Programming Interface (API) architecture (Google, 2018). This design was chosen for its predictability, scalability, and statelessness. Resources (e.g.,

participants, tasks, responses) were represented by nouns in the URL paths, and the standard HTTP method POST was used to denote actions upon these resources. To ensure statelessness, each request from the client contained all necessary information for the server to process it, with no session data stored on the server. The following table provides a general overview of each API endpoint's functionality.

Table 2. Endpoint descriptions

Endpoint	HTTP Method	Description
/api/authenticate	POST	Sends user's login name and password. Receives a validation success/unsuccess flag, user ID, username, and an indicator for first-time or returning user status
/api/goal-name	POST	Posts a user's new goal name. Receives a success or error message.
/api/respond-llm	POST	Posts user input for goal formation. Receives LLM-generated goal suggestions or feedback, along with success or error messages.
/api/submit-assessment	POST	Submits a user's numerical assessment mark. Receives a success or error message.
/api/respond-llm-assessor	POST	Posts user input for goal assessment. Receives LLM-generated assessment feedback, along with success or error messages.
/api/get-user-goal-name	POST	Fetches the user's latest goal name from the database. Receives the goal name or null, along with success or error messages.
/api/get-user-goal	POST	Fetches the user's latest detailed goal from the database. Receives the goal content or null, along with success or error messages.
/api/get-ordered-user-assessments	POST	Fetches a chronological list of a user's assessment content, the latest one for each unique day of submission. Receives these assessments or an empty object, along with success or error messages.

4.4 Data Persistence and Structure: Database Design

The foundation of the data analysis presented in this thesis is the dataset collected by the web application. The structure and integrity of this data were ensured through a relational database

design. The database schema serves as the formal "codebook" for the dataset, defining every variable and its relationship to others, thereby ensuring clarity and reproducibility in the analysis phase (Rob & Coronel, 1993). A relational database model, implemented using MySQL, was chosen for this project.

Key logical entities of the database schema

The logical structure of the database is designed around a set of key entities that represent the core objects and interactions within the application. This design prioritizes the organization of data by type of content and user interaction, allowing for direct capture of specific actions and their associated results. The key logical entities, directly corresponding to the database tables, are:

- Users, with their profile stored in the table “user_profile”
- Weekly textual progress assessments in the table “assessments”
- Textual short goal names in the table “goal_names”
- Textual detailed goal names in the table “goals”
- Weekly numerical progress assessments in the table “marks”

The **ER diagram** of the database is shown below.

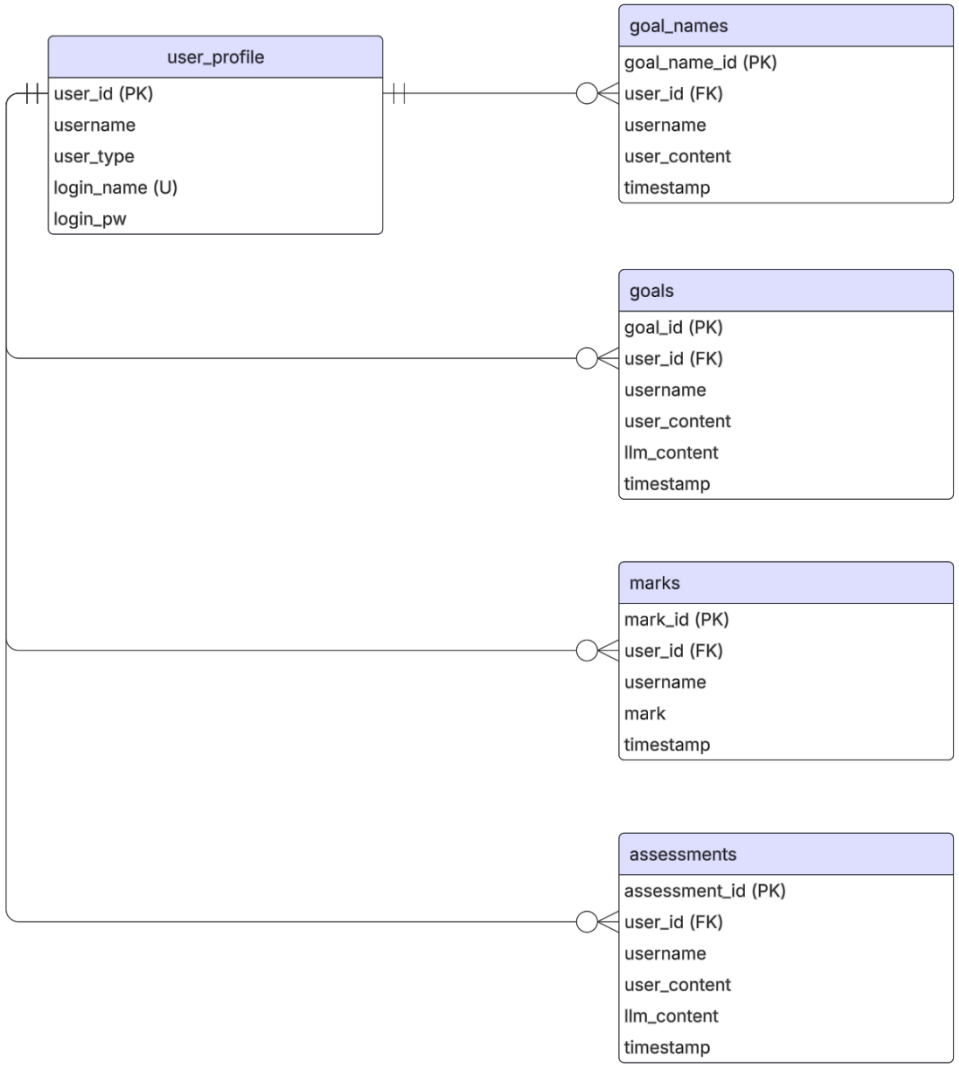


Figure 1. Physical ERD of the database

The figure above shows the ER diagram of the physical (as-built) schema; it reflects only the constraints enforced by the database. During the pilot, two additional application-level invariants were imposed by the business-logic layer (not by the database): (i) each user could record a goal name at most once, and (ii) usernames were treated as unique. As shown in the ER diagram, the username field is present in all five tables. While this design was a pragmatic choice for the pilot study to simplify data retrieval for analysis, it introduces data redundancy and violates principles of database normalization. For the project's current scale, this redundancy does not compromise application functionality. However, the business logic layer also enforced application-level

invariants, ensuring consistency despite the denormalized schema. A next iteration of the application would rectify this to enhance long-term data integrity and adhere to best practices.

Data access policy. The database was reachable only from the application server. Only the developer (author) and the supervisor held VM/DB admin rights; no pupils, teachers, or other researchers had database credentials, nor were there any scheduled ETL jobs or BI tools with write permissions. Consequently, all user data modifications occurred exclusively through the application's business-logic layer. This setup prevented race conditions and schema-bypassing writes, ensuring that integrity constraints not enforced by the database were upheld at the application layer.

4.5 Web Server Architecture

As a pilot project with an extremely low and infrequent traffic load, the web server architecture was designed with a focus on simplicity and resource efficiency. This made Flask's built-in development server, Werkzeug, an acceptable choice despite its limitations in a high-traffic environment.

The architecture comprises two primary components working in tandem: Nginx and the Flask application server.

- **Nginx (Reverse Proxy and SSL/TLS Terminator):** Nginx serves as the public-facing entry point for all incoming web traffic (HTTPS). It acts as a reverse proxy, intercepting client requests and forwarding them to the internal Flask application. Critically, Nginx performs SSL/TLS termination, decrypting incoming HTTPS requests and forwarding them to the Flask server over an unencrypted local connection. Conversely, it encrypts the responses received from Flask before sending them back to the user's browser over a secure channel.

- **Flask Development Server (Werkzeug):** The web application runs on Flask's integrated development server, which is responsible for receiving requests forwarded by Nginx. This component routes requests to the appropriate application logic and generates dynamic responses. Due to the minimal number of static files (e.g., one CSS and one JavaScript file per session), the Flask server was also used to serve these assets.

4.6 Deployment and Operational Infrastructure

The application was deployed on a virtual machine (VM) provisioned with Ubuntu 22.04 LTS, a standard, stable, and well-documented operating system provided by Tartu University's system administrators. This VM served as the host environment for a containerized architecture managed by Docker.

This approach was chosen to ensure a consistent and isolated environment for all application components. The entire web application was orchestrated through three separate Docker containers, which communicated with each other over a private Docker network:

- **Backend Container:** Hosted the Flask application and its business logic.
- **Database Container:** Contained the MySQL database, with a volume mounted to the host machine to ensure data persistence.
- **Nginx Container:** Acted as the reverse proxy, as described in section [insert section number for Web server architecture], directing traffic to the backend container.

The deployment workflow for application updates was executed through a series of methodical, largely manual steps, a pragmatic approach maintained throughout the pilot project's lifecycle. While this process was effective and reliable for the project’s limited scope, it is acknowledged that this method does not fully adhere to contemporary industry best practices for automation. A detailed, step-by-step description of this workflow is provided in Appendix 4.

5 Empirical Analysis

This section presents the details of data collection and the results of the experimental study examining the effects of LLM-based feedback on pupil inputs. The analysis began with a baseline similarity check to ensure group comparability, followed by comparisons of mean values between the test and control groups. To illustrate the nature of the inputs analysed, two example pupil–LLM interactions – one a rich multi-turn exchange and the other a brief single-input response – are provided in Appendix 3.

5.1 Data Collection

For the experiment, all 31 pupils at a single primary school in Tallinn were recruited. This school was selected due to an existing partnership between the school and the University of Tartu. Recruitment was conducted through teachers, who informed both pupils and their parents/guardians about the experiment and emphasized that participation was voluntary. No incentives or compensation were offered to either teachers or pupils, and no specific dropout prevention strategies were employed.

Pupils were randomly assigned to control and test groups using a random number generator to ensure unbiased allocation. The control group comprised 15 pupils, while the test group comprised 16 pupils.

The data was collected on the following dates:

1. **Goal Formation Session** (2025-05-05): Initial goal setting
2. **Assessment Session 1** (2025-05-12): First weekly progress assessment
3. **Assessment Session 2** (2025-05-19): Second weekly progress assessment
4. **Assessment Session 3** (2025-05-26): Third weekly progress assessment
5. **Assessment Session 4** (2025-06-02): Fourth weekly progress assessment
6. **Assessment Session 5** (2025-06-10): Fifth weekly progress assessment

Due to variable student attendance patterns, group composition differed across weekly sessions. As shown in Figure 2, attendance rates varied significantly: while over 80% of the test group

participated in all sessions, only slightly over 50% of the control group participated in all sessions. Since separate baseline checks were carried out for all groups, participation/absence patterns were not investigated any further.

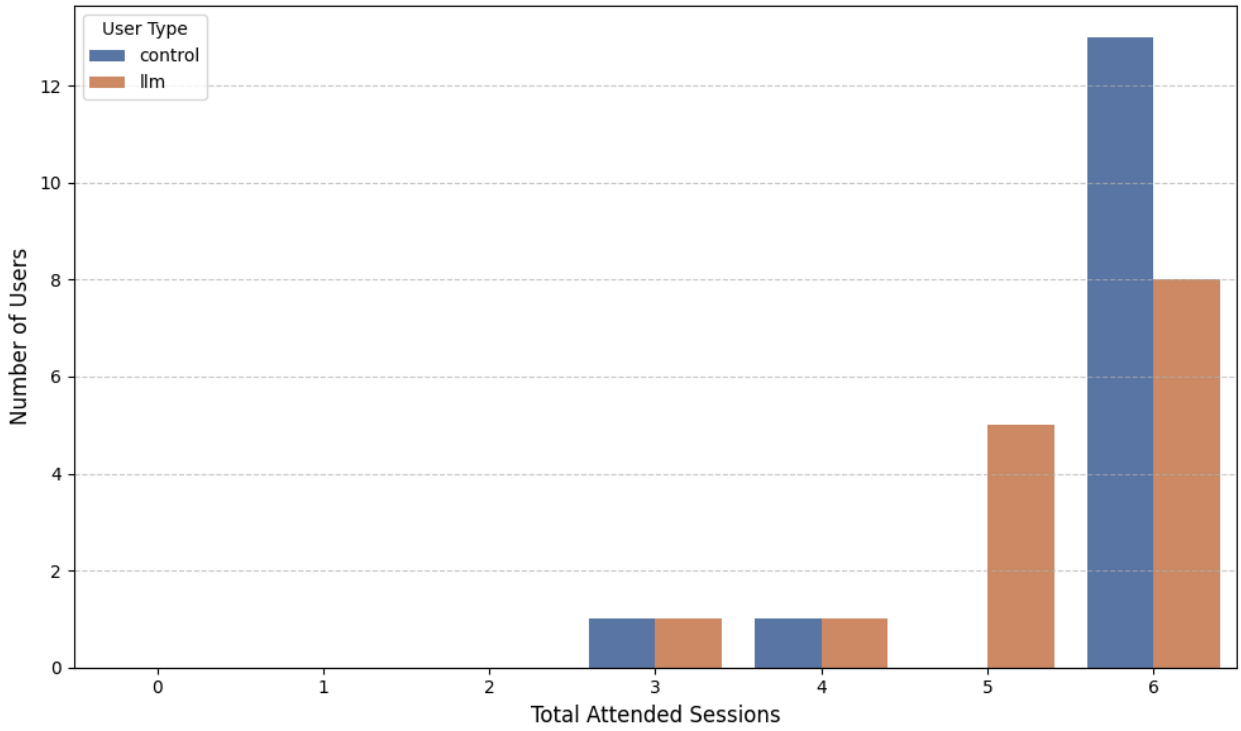


Figure 2. Distribution of number of attended sessions per user type

5.2 Data Preparation and Analysis

The analysis of the experimental data followed an iterative, two-stage methodology designed to accommodate the exploratory nature of the pilot study. This approach prioritized flexibility to allow for a detailed investigation of each specific research question.

- **Data Extraction and Initial Aggregation (SQL):** For each research question, a custom SQL query was developed and executed directly within the MySQL database environment. These queries were designed to aggregate raw input data and extract key metrics relevant to the analysis objective (e.g., calculating mean character counts or average progression

marks). The results of these queries were then manually extracted and prepared for the next stage of analysis.

- **Statistical Analysis and Visualization (Python):** The aggregated data was then ingested into a Python environment, typically using a Google Colab notebook, for further statistical analysis and visualization. Due to the small, manageable size of the datasets, this transfer was performed manually. The analysis relied on key Python libraries, including Pandas for data manipulation, SciPy for conducting statistical tests (e.g., t-tests), and Matplotlib, Seaborn for generating visualizations.

While the end-to-end workflow was a manual process, the core computations and statistical analyses were performed using automated tools such as SQL and Python libraries. This two-stage methodology was a pragmatic approach that allowed for a detailed investigation of the specific research questions of this pilot study.

An example of a code is given in Appendix 5.

5.3 Baseline Similarity Testing

Given the small group sizes (15 and 16 students respectively), the absence of even one or two students represents a 6-14% change in group composition. To address this concern, baseline similarity checks were conducted for each session by comparing pre-feedback input characteristics between treatment and control groups.

Table 3. Baseline similarity testing results

Session	Control Group Mean	LLM Group Mean	Difference	t-statistic	p-value
Goal Formation (2025-05-05)	63.37	40.60	-56%	-2.72	0.01*
Assessment 1 (2025-05-12)	71.29	78.30	9%	0.40	0.69
Assessment 2 (2025-05-19)	78.38	58.00	-35%	-1.27	0.22
Assessment 3 (2025-05-26)	70.64	73.43	4%	0.20	0.84

Assessment 4 (2025-06-02)	68.57	77.77	12%	0.60	0.56
Assessment 5 (2025-06-10)	76.00	77.43	2%	0.07	0.95

*Statistically significant difference ($p < 0.05$; Bonferroni correction not applied as this was a single session analysis)

There was a statistically significant baseline difference in the goal formation session ($p=0.01$): control group members posted significantly longer inputs than test group members. However, all assessment sessions showed non-significant baseline differences ($p > 0.01$, with Bonferroni correction applied across the 5 assessment sessions). This means that groups were comparable for all assessment session analyses.

5.4 Research Statement Exploration

The research statements primarily examined the relationship between input length and reflection quality, as well as changes in reflection depth resulting from feedback.

5.4.1 Input length and reflection depth: qualitative insights

A qualitative review of pupils' inputs suggests that longer responses often contained more elaborated reasoning, explicit self-reflection, and richer detail about personal learning strategies. Shorter responses tended to be more superficial, frequently limited to single statements without explanation or justification. However, this relationship was not uniform: some long responses contained filler text or tangential remarks, while some short responses were concise yet meaningful.

The table below presents three inputs with low reflective depth and three inputs with high reflective depth.[2] The pattern is visible: less sophisticated inputs tend to be shorter, and more sophisticated inputs tend to be longer.

[2] In the subsequent tables, translations into English were prepared by the author and edited with the assistance of a large language model. The translations preserve original typos, capitalisation, punctuation errors, and other inaccuracies.

Table 4. Input length and reflection depth

Short and superficial inputs	Longer and deeper inputs
Läks halvasti (It went badly)	Ma lugesin teksti aeglaselt ja keskendusin sellele.Kui ei saanud midagi aru lugesin uuesti ja korralikumalt.See eesmärk aitas mul arusaada rohkem tekstistTundsin ennast suurepäraselt. (I read the text slowly and concentrated on it. If I didn't understand something, I read it again more thoroughly. This goal helped me understand the text better. I felt great.)
saada kõik iga nädal järje peale (to catch up with everything each week)	Mul on korras kõik peale matemaatika, milles on vaja lõpetada üks osa ja teine teha, ajalugu ja kirjandus. Matemaatika teen täna kodus ja ajaloo teen homme. Kui jõuan siis teen homme kirjanduse ka kodus. (I have all in order except mathematics, in which need to finish one part and do another, history and literature. Mathematics do today at home and history do tomorrow. If I manage then do tomorrow literature also at home)
mul on õppimisega pekkis (I'm totally messed up with studying)	Enamasti õppisin sisulielt ja s in kõigidest teemadest aru, mõnikord juhtus, et ei saanud aru, aga siis ma küsisin. Mul läks vahepeal mõte mujale ning siis küsisin (Mostly I studied with focus and s in understood all topics, sometimes it happened that did not understand, but then I asked. My mind went elsewhere in between and then I asked)
väga halb (very bad)	Enamasti sain kõigest aru ja kui ei saanud siis kohe küsisin. Mul aitas aru saada, kui keegi selgitas mulle ära. Peale seda, kui aru sain oli hea tunne. Mul aiatas ka aru saada skeemide vaatamine. Mulle tõid abi skeemid mis olid lihtsad ja värvidega eristatud. (Mostly understood everything and if did not then immediately asked. Helped me to understand when someone explained to me. After that, when understood was good feeling. Helped me also to understand looking at schemes. Brought me help schemes which were simple and distinguished with colours)
jäi teha palju (there's still a lot to do)	Harjutasin ilusalt kirjutamist kodus,aga tihti unustasin ära kui tegin ülessandeid. Unustasin

	sellel hetkel kui ma õppisin koolis ja oli kiirem, sellepärast kirjutasin lohakalt. Kui mul on kiire siis ma pean lihtsalt mõtlema ,et vahet ei ole kas ma opin kodus voi koolis. Lihtsalt ei tohi teha asju lohakalt. (Practiced writing nicely at home,but often forgot when did tasks. Forgot at that moment when I studied at school and was faster, because of that wrote sloppily. When I am in a hurry then I just have to think ,that it doesn't matter whether I study at home or at school. Just must not do things sloppily)

The short responses above lack detail about what happened and what the pupil plans to do next. They also lack analytical insight into why the situation is as it is or why certain steps should be taken – elements that are central to meaningful reflection.

By contrast, the longer responses describe the pupil's internal processes – thoughts, strategies, and feelings related to studying – and often include cause-and-effect reasoning absent from shorter inputs.

There were also inputs that ranked relatively high in reflection quality despite being brief in length. Table 5. below presents examples of such cases.

Table 5. Shorter inputs with notable reflection depth

ma õppisin mõttega ega kiirustanud olen enda üle uhke et täitsin eesmärgi (I studied with focus and without rushing; I am proud of myself for achieving the goal)
Mulle meeldis korralikult tegutseda ja oli hea vaadatata korrektset tööd. (I enjoyed working properly, and it was nice to see neat work)
ma sain mitte eriti hästi hakama kuna ma ei saanud keskenduda (I didn't manage very well because I couldn't concentrate)
See et ma loen keskendudes ja ei mõtle teistele asjadele (That I read with focus and do not think about other things)
Täitsin eesmärgi tunnen et eesmärk on aidanud mul lugedes keskenduda olen rahul. (I fulfilled the goal and feel that having the goal has helped me focus on reading. I am satisfied.)

These inputs demonstrate the use of abstract and self-analytical concepts to describe the pupil's state of mind and learning process. Therefore, their quality of reflection can be considered notable – even if not the highest – despite their relatively short length.

Two conclusions emerge from these observations:

- There appears to be a positive connection between input length and reflection quality.
- The connection is not definitive—input length, while potentially indicative of depth, cannot serve as a reliable standalone measure of reflection quality.

These findings serve as sufficient justification for further exploration of input length patterns.

Incomplete inputs

Some inputs were evidently incomplete or abruptly cut off, suggesting they were submitted unintentionally or before the pupil had finished typing. Examples are shown below:

Table 6. Incomplete inputs

Ma (I)
poool [no translation; not a meaningful word]
Nädala alguses ma ei (In the beginning of the week I didn't)
Ma sain kõik mis vaja valmis ja koduõppep (I got everything I needed done and home stu) [incomplete]
et ma oleks sõpradega rohkem (that I would be more with friends) [incomplete]

Such entries were likely the result of accidental clicks on the "Send" button or premature submission. While these inputs were retained in the dataset to preserve the integrity of the raw data and to better understand its nature, they introduce noise into analyses based purely on length or textual similarity.

5.4.2 Research Statement1: Goal Formation Length Patterns

The test group showed a tendency toward longer goal statements compared to the control group. However, only one post-treatment input was collected in the control group, making meaningful

statistical comparison impossible. The apparent pattern cannot be attributed to LLM feedback effects, as the comparison lacks sufficient control group data for valid inference.

5.4.3 Research Statement 2: Weekly Progress Marks Patterns

The test group showed a tendency toward higher self-reported weekly progression marks compared to the control group (Table 7). This pattern was observed in 4 of 5 sessions, with the test group reporting marks that were, on average, 8.4% higher (2.46 vs 2.27 on the 3-point scale).

Even though the directional pattern was consistent – the test group reported higher progression marks in 4 of 5 sessions – there were no statistically significant differences after Bonferroni correction (α=0.01).

Table 7. Weekly progress marks, 3-point Likert scale

Session	Control Group Mean	LLM Group Mean	Difference	t-statistic	p-value
Assessment 1 (2025-05-12)	2.20	2.40	8%	0.92	0.37
Assessment 2 (2025-05-19)	2.06	2.50	18%	1.80	0.08
Assessment 3 (2025-05-26)	2.33	2.13	-9%	-0.80	0.43
Assessment 4 (2025-06-02)	2.24	2.60	14%	1.60	0.12
Assessment 5 (2025-06-10)	2.53	2.65	5%	0.63	0.54

This exploratory pattern suggests that exposure to LLM feedback may be associated with slightly more positive self-assessments of goal progression, though the lack of statistical significance indicates this should be interpreted cautiously. Moreover, attributing this difference to LLM feedback mechanisms is problematic because users in both groups click the assessment button before receiving any LLM feedback, suggesting other factors may explain the observed pattern.

5.4.4 Research Statement 3: Weekly Progress Assessment Patterns

The test group showed a consistent tendency toward longer weekly progress assessments compared to the control group across most sessions (Table 8). This pattern was observed in 4 of 5 sessions, with the test group producing inputs that were, on average, 38.8% longer.

When examining individual sessions with Bonferroni-corrected significance levels (α=0.01), Session 3 showed a statistically significant difference (p=0.0098). While Session 5 approached significance (p=0.0106), it did not meet the corrected threshold.

This exploratory finding suggests that exposure to LLM feedback may be associated with more elaborate progress assessments, though the pattern varies across sessions and requires further investigation to understand the underlying mechanisms.

Table 8. Post-treatment input length analysis

Session	Control Group Mean (chars)	LLM Group Mean (chars)	Difference	t-statistic	p-value
Assessment 1 (2025-05-12)	113.75	148.25	23%	0.73	0.50
Assessment 2 (2025-05-19)	68.50	80.50	15%	0.35	0.79
Assessment 3 (2025-05-26)	72.00	127.50	44%	3.14	0.01* (0.0098)
Assessment 4 (2025-06-02)	90.50	86.87	-4%	-0.18	0.86
Assessment 5 (2025-06-10)	77.75	143.50	46%	2.85	0.01 (0.0106)

5.4.5 Research Statement 4: Input Similarity Distributions

Two similarity measures were calculated: normalized Damerau-Levenshtein similarity (1-Damerau-Levenshtein distance) and Cosine similarity. Kolmogorov-Smirnov tests were conducted to test if the distributions were drawn from the same underlying distributions (the similarity score distributions for both groups come from the same underlying distribution) or not (there is a difference).

Results are presented in the Figure 3 below. This shows that neither similarity measure showed differences between control and test group. It means that there were neither significant differences in correcting typos and punctuation nor in changes in meanings between the test and control groups.

Both metrics show that there was a considerable number of identical pairs. Indeed, examination of the inputs showed that in many cases pupils had submitted the same input several times, without making any change. The percentage of such unchanged inputs was lower in the test group. This finding aligns with other findings of the study suggesting that LLM-feedback had an immediate impact. However, the effect was not statistically significant. In the case of statistical significance i.e. if the distributions were different, it would have offered only remote support to the statement because this pattern was only a part of the distribution.

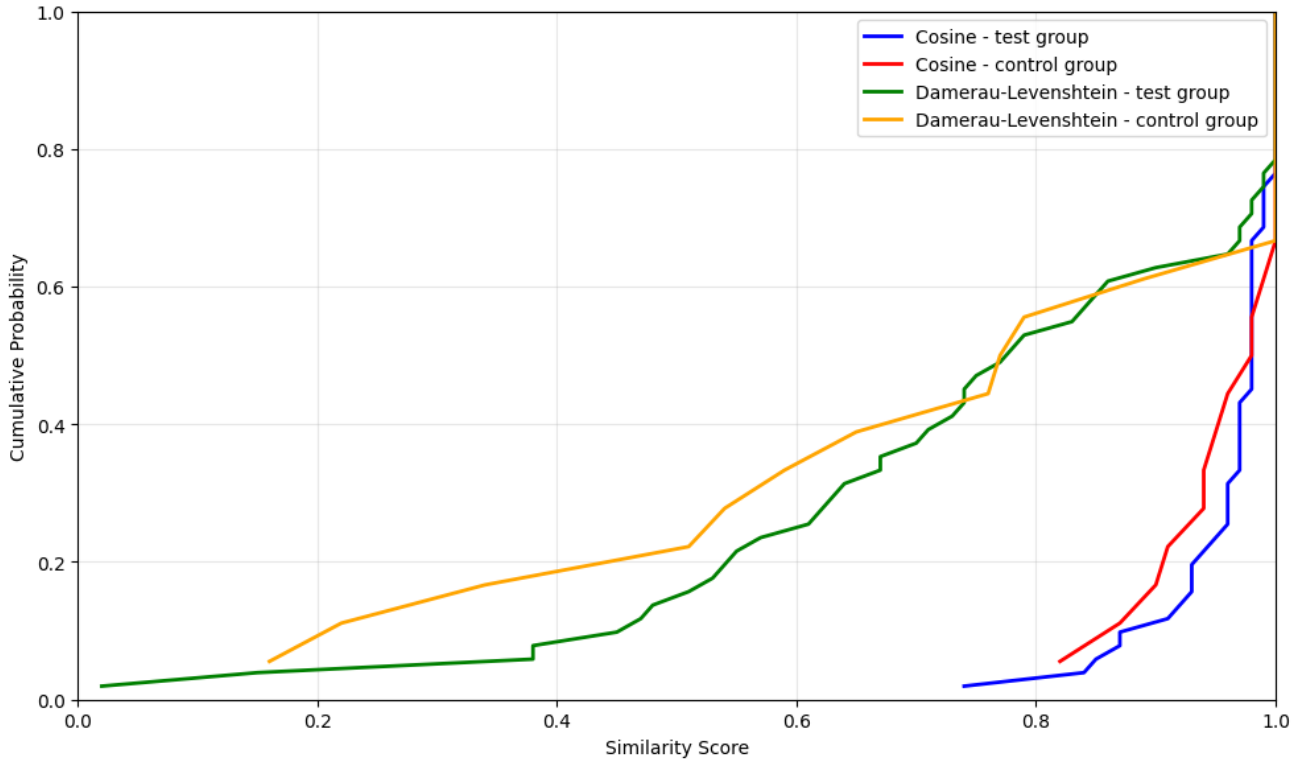


Figure 3. Cumulative distribution functions of the similarity measures

The table below presents the summary statistics for the distributions, helping to understand the significance of differences between the groups.

Table 9. Input similarity distribution analysis

Similarity measure	Test Statistic (D)	p-value
Cosine similarity	0.190	0.65
1-Damerau-Levenshtein distance	0.154	0.86

Results in the table above show there are no statistically significant differences in similarity score distributions between groups. In other words, the finding shows that the distributions of change magnitudes are similar between test and control groups – meaning both groups show comparable patterns in how much they modify their inputs, regardless of whether they received LLM feedback or not.

This finding, however, says nothing about whether the two similarity measures produce concordant values or rankings for the same input pairs. That would be a separate analysis of inter-measure reliability which was not conducted in this exploratory phase. Such inter-measure reliability analysis would be more appropriate for future confirmatory studies that aim to establish standardized measurement protocols for LLM feedback impact assessment. For this study, these two measures were chosen to capture different aspects of text similarity (lexical vs. semantic) to gain a more comprehensive understanding of the input changes.

5.5 Interview Findings

The interviews with pupils revealed that they perceived the LLM-generated feedback as motivating and helpful for reflecting on their goals and progress. Specifically, pupils reported that the feedback prompted them to critically examine the wording of their goals and to revise them accordingly. It also supported their planning for the upcoming week. Overall, the feedback was perceived as useful. Pupils in the control group – who did not receive LLM feedback – expressed that they would have liked to receive such feedback as well.

Importantly, no pupil mentioned negative user experiences or system flaws that would have prevented them from using the web application.

The interviews with teachers indicated that they believed the webapp increased pupils' motivation and interest in setting personal goals and monitoring their progress. Teachers also observed that the feedback helped pupils formulate more concrete goals and supported the development of reflective thinking skills and habits. According to them, the webapp had two distinct effects: on an operational or technical level, it enhanced reflection skills; on an emotional or metacognitive level, the personalized feedback gave pupils the sense that they were not alone with their goals, but that someone else was engaged and supportive.

The teachers emphasized that the app should not be seen as a replacement for a human teacher. One reason for this is that not all pupils understand or trust AI-based systems to the extent required for them to use such tools effectively on their own.

Similar to the pupils, the teachers did not identify any major flaws or system limitations that would have prevented its use.

In sum, the interview findings align well with the results obtained from the analysis of the webapp-generated data.

5.6 Summary of Findings

This pilot produced a coherent, though not statistically conclusive, pattern of results. In multiple indicators, the test group showed tendencies toward longer responses, higher progress marks, and more frequent revisions, while interview feedback from both pupils and teachers described the LLM-generated feedback as helpful for thinking more deeply about goals and progress. These patterns were consistent in direction but modest in magnitude and did not reach statistical significance, a result influenced by limited sample size, few inputs per participant, and high variability within groups.

Beyond these headline results, the study generated operational insights that are critical for planning a larger-scale investigation. First, it established realistic expectations for data volume and variability, enabling more precise sample size calculations. Second, it demonstrated the feasibility and limitations of readily available metrics such as character count and similarity scores. Finally, it surfaced practical considerations – from interface behaviour to text quality – that directly inform refinements to the application's architecture, data pipeline, and analytic approach.

6 Limitations of the Study and Recommendations

This study was a pilot study conducted to inform a possible larger study that focuses on establishing the effects of using LLM for giving personalized feedback to pupils in their pursuit of formulating and progressing toward school-related goals. This section reviews lessons learned from this study and formulates recommendations for the new study. These are presented in four sections: validity-related limitations and recommendations, experiment design and sample size related limitations and recommendations, technology stack related limitations and recommendations, and limitations of practical considerations and related recommendations covering training, standardization, ethics, and monitoring.

6.1 Validity

Construct validity refers to the extent to which a study accurately measures the theoretical concepts it intends to investigate. In this pilot, the impact of LLM-generated feedback was explored using two readily available indicators: average input length and textual similarity between consecutive user inputs. While these do not capture the full depth of constructs such as metacognitive engagement or reflection quality, they did reveal patterns of engagement that may be meaningful. For example, longer responses and certain changes between pre- and post-feedback inputs were more common in the test group, suggesting that such metrics can provide initial signals of interaction effects. However, their interpretive power is limited without complementary measures – such as qualitative coding, semantic analysis, or rubric-based human evaluation – which could better align future analyses with the underlying theoretical constructs. In this sense, the pilot supports the feasibility of using simple metrics as a first step, while pointing to methodological refinements needed for richer interpretation.

6.1.1 Preprocessing as a Prerequisite for Construct Validity

While misspellings, inconsistent capitalisation, and incomplete or abruptly terminated inputs may appear to be mere “noise,” their impact can be systematic rather than random. Such errors can distort calculated input lengths, obscure semantic meaning, and mislead both human and automated assessments of reflection quality. If unaddressed, they reduce construct validity and can bias results in consistent ways rather than simply adding random variation. Therefore, a methodologically solid approach to preprocessing the raw inputs needs to be developed before

substantive analysis methods can be applied. At this point, the most promising avenue appears to be merging large language model capabilities with algorithmic approaches. Crucially, this process should take 10–13-year-olds' knowledge and capabilities as the starting point—recognising some errors as natural while still filtering out clear issues such as single words, meaningless fragments, or accidental keystrokes.

External validity refers to the extent to which the findings of a study can be generalized beyond the specific conditions of the experiment. It evaluates whether the results are applicable to other populations, settings, or times. Factors influencing external validity include the sampling method, the representativeness of the participants, and the ecological validity of the study's context.

External validity of this study is limited by several factors. The sample that was obtained from one school that employs distinctive teaching methods probably does not represent the broader population of 10-13 years old primary school pupils. Therefore, the sample is likely to introduce a systematic bias. Also, this small sample is susceptible to random deviations. This limitation needs to be addressed in the full study by selecting more common or mainstream schools and enlarging the sample. Additionally, the study focuses on sessions, examining short-term immediate effects. This excludes potential long-term outcomes that may arise from using such web application on a longer-term basis. However, using AI-enhanced platforms on a longer-term basis is a commonplace practice that is not likely to change. The full-scale study should address this usage pattern and attempt to capture long-term individual level changes.

Internal validity refers to the degree to which the results of a study can be attributed to the interventions or variables being tested, rather than other factors. This is influenced by proper experimental controls, such as random assignment, standardized procedures, and baseline checks to minimize the effects of confounding variables.

In this study, several measures were taken to ensure internal validity. Random assignment into treatment and control group was implemented. Standardized environment, web application and participation procedures for both groups ensured consistent treatment delivery. Baseline checks for each session data controlled composition effects due to variable attendance. There were no

major flaws in this respect. However, it is worth emphasizing that all these measures should be maintained in future studies.

6.2 Experiment Design

This pilot used a cross-sectional design, comparing single-session data rather than tracking individual pupils over time. While quick to implement, this limits the ability to see how goal-setting and reflection practices evolve with repeated use. In a practical school context, tracking individual progress would provide richer insights and make feedback more relevant and motivating.

A longitudinal approach—following the same pupils across multiple sessions—would offer two key benefits. First, it would reveal genuine changes in goal-setting quality and reflective depth over time, rather than relying on group averages. Second, it would help teachers identify pupils needing targeted support, even if class averages remain unchanged. In addition to its theoretical value, a longitudinal design can also achieve the same statistical power with fewer participants, as within-subject comparisons reduce error variance.

The control group setup also warrants adjustment. In this pilot, the control group received no feedback, which likely contributed to disengagement and higher dropout. In classrooms, pupils almost always receive some form of feedback, even if generic. To better reflect real practice and sustain engagement, the control group could instead receive short, pre-written prompts randomly drawn from a small library. This preserves the experimental contrast between personalized and generic feedback while keeping all participants engaged.

While generic control feedback may narrow the difference between groups, the trade-off could be worthwhile: slightly weaker contrasts in the data may be acceptable if it yields higher participation, better user experience, and findings that better represent real classroom use.

6.3 Automation of Data Analysis

While the two-stage manual analysis workflow was a pragmatic and effective choice for this pilot study, a next iteration of the application would require a fully automated analysis pipeline. The manual process of extracting data via SQL and performing subsequent analysis in a separate environment would not be scalable to a larger user base or more frequent data collection. To ensure

reproducibility, reliability, and efficiency, future development would need to integrate the data extraction and analysis into a single, automated script. This would programmatically connect to the database, perform all necessary statistical analyses, and generate reports without manual intervention.

Automating the analysis presents specific challenges related to the quality of pupil-generated text. As documented in Table 6. Incomplete inputs, some entries are abruptly cut off, likely due to premature submission. In addition, as shown in Table 4 in Table 5, many inputs contain typos, misspellings, and non-standard sentence structures. These issues highlight a key limitation in using unedited text for automated analysis: validity can be compromised not only by semantic ambiguity but also by input completeness and surface-level errors. Addressing these systematically through preprocessing is therefore a prerequisite for automation (see section 6.1.1).

6.4 Sample Size

Based on the data and results obtained in this study and the recommended longitudinal research design, sample sizes were calculated for various conditions that might occur in a full-scale study. The pilot study conditions would require 171 participants for single measurement or 89 for repeated measurements. Under more challenging conditions with reduced effect size, these numbers would increase to 754 and 392 respectively. More detailed calculations and scenario explanations are provided in Appendix 7.

6.5 Technology Stack

The current technology stack, comprising a React-based UI, Flask business logic server, MySQL database, and Nginx reverse proxy, performed adequately under the load generated during this study with 31 total users and up to 20 concurrent users per session. However, scaling to approximately 170 participants would require significant upgrades to handle increased concurrent requests.

The most pressing issue involves potential concurrency problems at two levels. First, the Flask backend currently handles requests sequentially and would require transition to an asynchronous framework such as FastAPI or deployment behind a scalable server to support concurrent request handling.

Second, LLM API interactions present a critical bottleneck, as the current system lacks rate-limiting and queuing mechanisms. Concurrent user submissions risk exceeding API rate limits, leading to service degradation and user attrition. Implementing multiple API keys and proper queuing mechanisms would be necessary.

Additional improvements would include implementing a separate webserver for static files, enhancing system monitoring and logging, implementing security measures against web vulnerabilities, and improving database integrity through proper password hashing and schema normalization.

6.6 Recommendations Based on Interviews

The interviews with pupils revealed several areas for improvement in user interface and user experience. Pupils expected a more engaging interface with brighter colors and the ability to view goals from previous weeks to track progress over time. They also expected feedback to reference previous interactions, creating continuity between sessions for a more personalized experience. Notably, all pupils expected to receive feedback, with those in the control group expressing dissatisfaction with its absence.

The interviews with teachers highlighted several important considerations for future iterations. Teachers emphasized the need for basic training to explain the application's purpose and usage, warning that without proper guidance, pupils would be unlikely to engage effectively. They also cautioned against making feedback appear too anthropomorphic, as overly human-like responses might not be taken seriously by pupils, thereby undermining the potential benefits.

Teachers suggested two additional features: a dashboard allowing them to view all pupils' inputs at a glance, and functionality to flag pupils who may need additional support in goal formulation or progress evaluation. Finally, they recommended integrating webapp access with existing school accounts to eliminate the need for additional login credentials.

7 Conclusion

The main research question of this pilot study asked: **Can readily available indicators provide meaningful insights into how pupils engage with LLM-supported feedback?** The study's four research statements explored specific, readily available metrics as potential indicators of such effects.

Research Statement 1: Goal Formation Length Patterns

Test group pupils tended to produce longer goal statements than the control group. However, due to a lack of post-treatment inputs from the control group in the goal-formation session, statistical comparison was not feasible. This highlights a measurement challenge: without sufficient comparable data points, the potential effect of LLM feedback on goal formulation length cannot be robustly assessed.

Research Statement 2: Weekly Progress Marks Patterns

The test group reported slightly higher self-assessed weekly progress marks in most sessions, but differences were not statistically significant and could not be causally linked to LLM feedback. Marks were recorded before feedback was given, indicating that any observed differences may be driven by other factors. This underscores a design limitation: outcome variables must be temporally aligned with the treatment to enable valid causal interpretation.

Research Statement 3: Weekly Progress Assessment Patterns

Across most sessions, the test group produced substantially longer written progress assessments, with one session showing a statistically significant difference and another approaching significance after Bonferroni correction. While suggestive of more elaborate reflection following LLM feedback, the inconsistency across sessions and small sample size prevent definitive conclusions.

Research Statement 4: Input Similarity Distributions

Lexical and semantic similarity measures showed no statistically significant differences between groups. While the proportion of unchanged inputs was lower in the test group, the overall difference was not significant.

Taken together, the results indicate that simple, scalable metrics – such as input length and changes in textual similarity – can reveal engagement patterns consistent with an effect of LLM-based feedback on pupils' goal-setting processes, even when those goals are extra-curricular rather than academic. These patterns emerged despite the small sample size and high control-group attrition, both of which limited statistical power. The findings suggest that readily available indicators can be a practical starting point for monitoring pupil engagement in real-world settings, particularly where more resource-intensive methods are impractical. At the same time, the study underscores the importance of refining measurement strategies, improving experimental design (e.g., longitudinal structure with non-personalized control feedback), expanding sample size, and incorporating preprocessing and more advanced natural language processing to enhance both validity and interpretive depth.

8 References

Aparna, K.S. (2024). Top 10 UI/UX design principles for creating child-friendly interfaces. [Blog post]. *aufaitUX*. https://www.aufaitux.com/blog/ui-ux-designing-for-children/

Barratt, H. & Kirwan, M. (2009). Shantikumar, S. (2018). Clustered data - effects on sample size and approaches to analysis. *HealthKnowledge.* https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1a-epidemiology/clustered-data

GeeksforGeeks. (2025) Client-server architecture - system design. [Blog post]. *GeeksforGeeks*. https://www.geeksforgeeks.org/system-design/client-server-architecture-system-design/

Georghiades, P. (2004). From the general to the situated: three decades of metacognition. *International Journal of Science Education, 26*(3), 365-383, DOI: 10.1080/0950069032000119401, https://www1.wellesley.edu/sites/default/files/assets/departments/pltc/files/faculty/georghiades_2004_decades_of_metacognition.pdf

Google. (2018). Web API design: the missing link best practices for crafting interfaces that developers love. *apigee*. apigee-web-api-design-the-missing-link-ebook.pdf

Guo, Q., Zhen, J., Wu, W., He, Y. & Qiao, C. (2025). Can students make STEM progress with the large language models (LLMs)? An empirical study of LLMs integration within middle school science and engineering practice. *Journal of Educational Computing Research*, *63*(2). https://journals.sagepub.com/doi/full/10.1177/07356331241312365

Higgins, J. P. T., & Green, S. (Eds.). (2011). 16.3.4 Approximate analyses of cluster-randomized trials: effective sample sizes. *Cochrane handbook for systematic reviews of interventions (Version 5.1.0).* The Cochrane Collaboration. https://handbook-5-1.cochrane.org/chapter_16/16_3_4_approximate_analyses_of_cluster_randomized_trials_for_a.htm

Kazemitabaar, M., Hou, X., Henley, A., Ericson, B.J. Weintrop, D. & Grossman, T. (2023). How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. Accepted for publication in the *Proceedings of the 2023 ACM Koli Calling International Conference on Computing Education Research*. https://arxiv.org/abs/2309.14049

Kersna, L., Laak, K.-J., Lepp, L., & Pedaste, M. (2025). Supporting Self-Regulated Learning in Primary Education: Using Written Learning Guides in the Lessons. *Education Sciences*, *15*(1), 60. https://doi.org/10.3390/educsci15010060

Kurt, S. (2023). Self-regulated learning: what it is, why it is important and strategies for implementing it. [Blog post]. *Educational Technology*. https://educationaltechnology.net/self-regulated-learning-what-it-is-why-it-is-important-and-strategies-for-implementing-it/

Li, P.H., Lee, H.Y., Lin, C.J., Wang, W.C. & Huang, Y.M. (2025). InquiryGPT: augmenting ChatGPT for enhancing inquiry-based learning in STEM education. *Journal of Educational Computing Research, 62*(8). https://journals.sagepub.com/doi/10.1177/07356331241289824

Lin, Y., Yang, E. F. Y., Wu, J. S., Yeh, C. Y. C., Liao, C. Y., & Chan, T. W. (2024). Enhancing students' authentic mathematical problem-solving skills and confidence through error analysis of GPT-4 solutions. *Research and Practice in Technology Enhanced Learning. 20*:034. DOI:10.58459/rptel.2025.20034

Martins Van Jaarsveld, G., Wong, J., Baars, M., Specht, M. & Paas, F. (2025). Scaling goal-setting interventions in higher education using a conversational agent: examining the effectiveness of guidance and adaptive feedback. 15th *International Conference on Learning Analytics and Knowledge, LAK 2025* (pp. 328-338). Association for Computing Machinery, Inc. https://doi.org/10.1145/3706468.3706510

Park, I., Winglee, M., Clark, J., Rust, K., Sedlak, A. & Morganstein, D. (2003). Design effects and survey planning. *2003 Joint Statistical Meetings - Section on Survey Research Methods.*

Parker, J. (2023). What is a three tiered client server architecture? [Blog post]. *Architecture*. https://www.architecturemaker.com/what-is-a-three-tiered-client-server-architecture/

Ramalho, L. (2022). Fluent Python. Clear, concise, and effective programming. 2nd Edition. O'Reilly.

Rob, P. & Coronel, C. (1993). *Database systems. Design, implementation, and management*. Wadsworth Publishing Company.

Shrout, P.E. & Fleiss, J. L., (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull., 86*: 420-428. https://www.aliquote.org/cours/2012_biomed/biblio/Shrout1979.pdf

Shute, V.J. (2008). Focus on Formative Feedback. *Review of Educational Research, 78*(1). DOI: 10.3102/0034654307313795

Ux&You (2023). UX and UI recommendations when designing for Kids. *Medium*. https://medium.com/@uxandyouti/ux-and-ui-recommendations-when-designing-for-kids-abe3769e5b2d

Veenman, M.V.J., Van Hout-Wolters, B.H.A.M & Afflerbach, P. (2006). Metacognition and learning: conceptual and methodological considerations. *Metacognition Learning 1*: 3–14. DOI 10.1007/s11409-006-6893-0

Wang, M., Zhang, D., Zhu, J., & Gu, H. (2025). Effects of incorporating a large language model-based adaptive mechanism into contextual games on students' academic performance, flow experience, cognitive load and behavioral patterns. *Journal of Educational Computing Research, 63*(3), 662-694. https://doi.org/10.1177/07356331251321719 (Original work published 2025)

Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S. & Wen, Q. (2024). *Large language models for education: A survey and outlook*. arXiv. https://arxiv.org/pdf/2403.18105

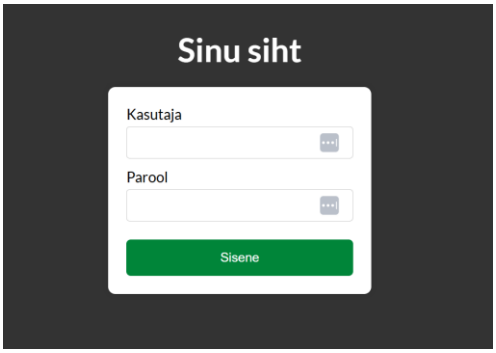
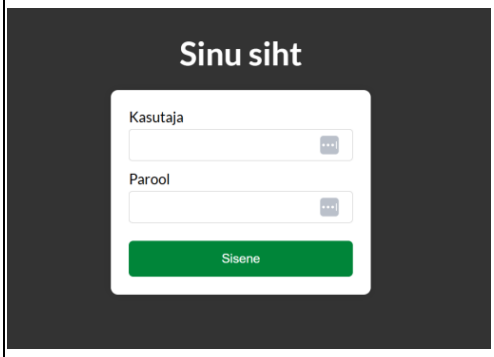
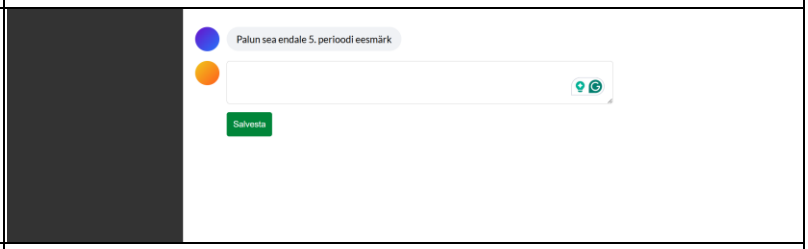
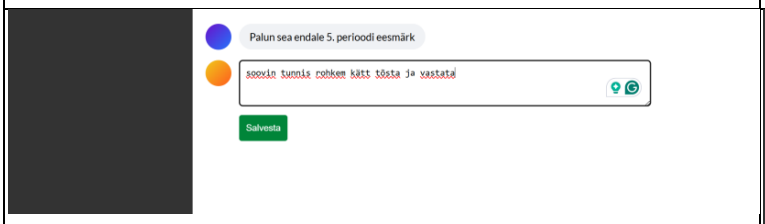
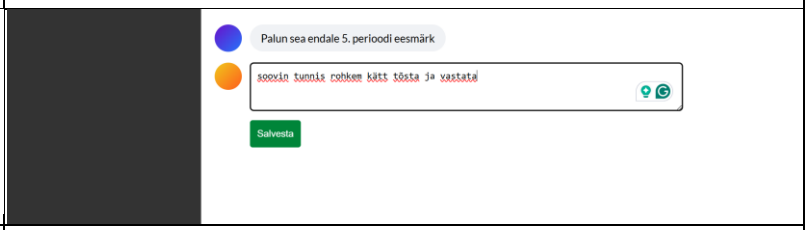
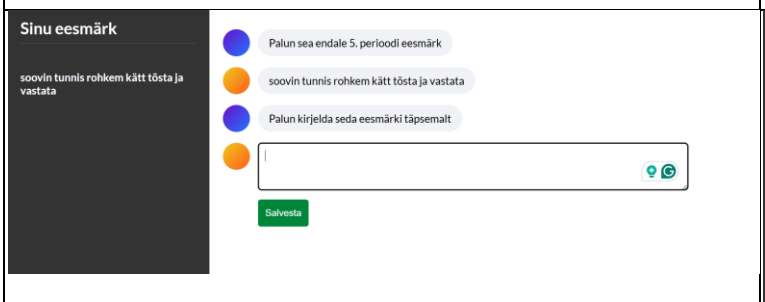
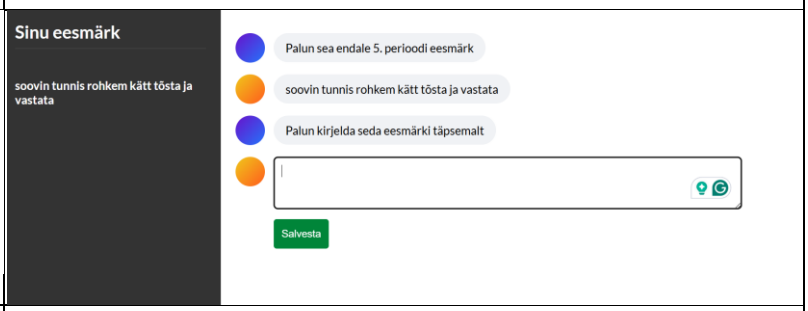
Zimmermann, B.J. (2004). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice 41*(2):64-70. DOI:10.1207/s15430421tip4102_2

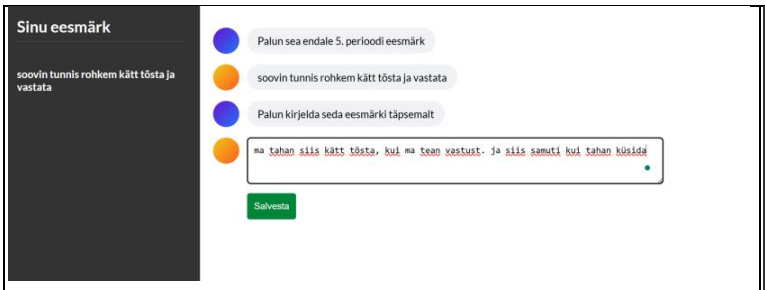
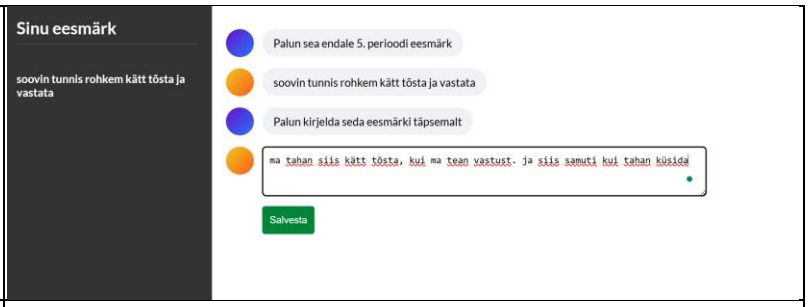
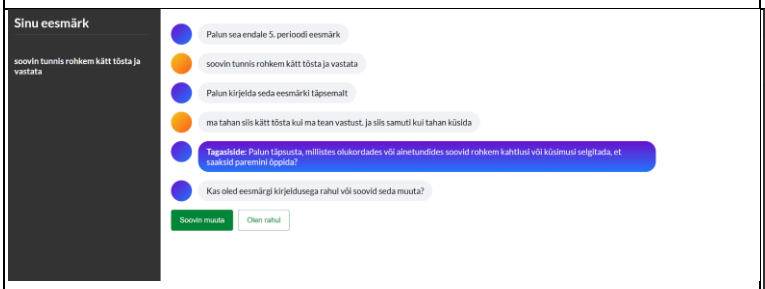
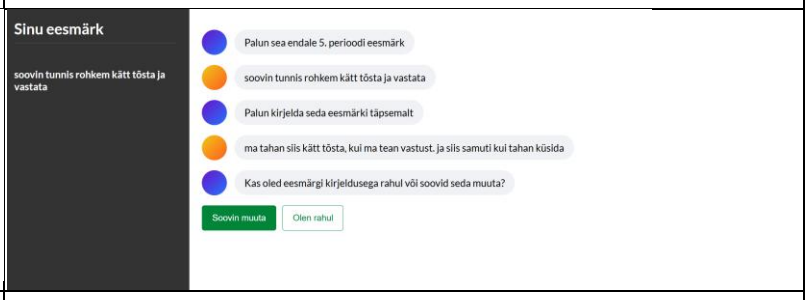
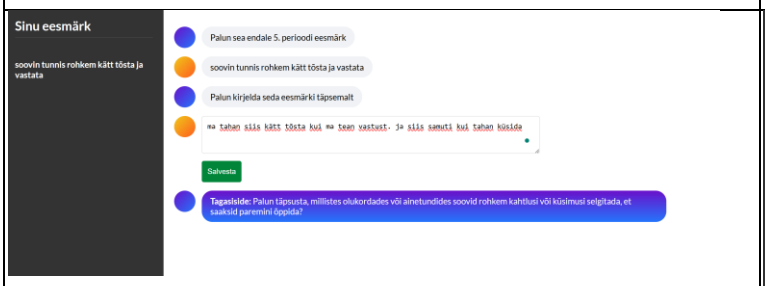
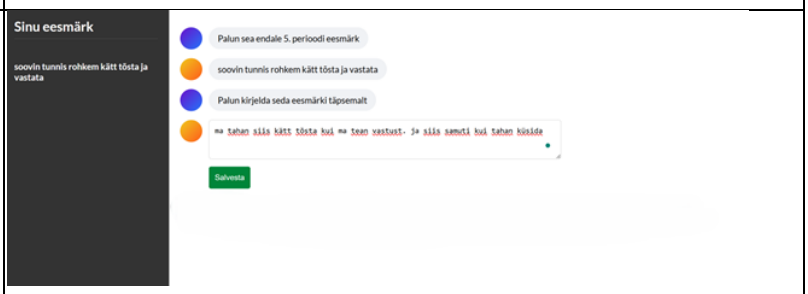
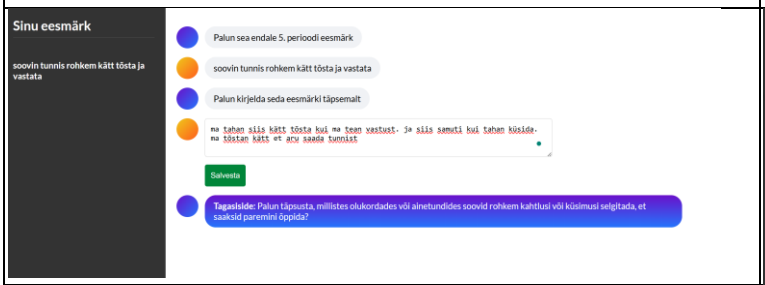
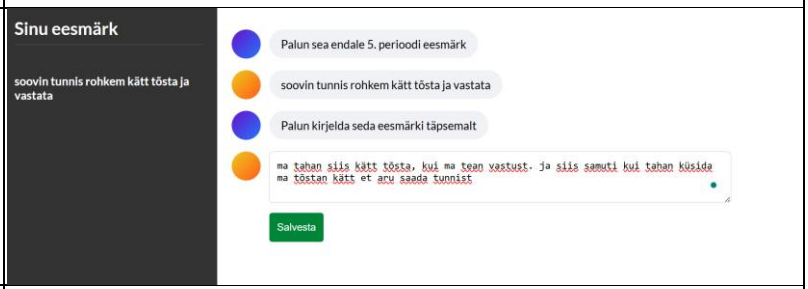
9 Appendices

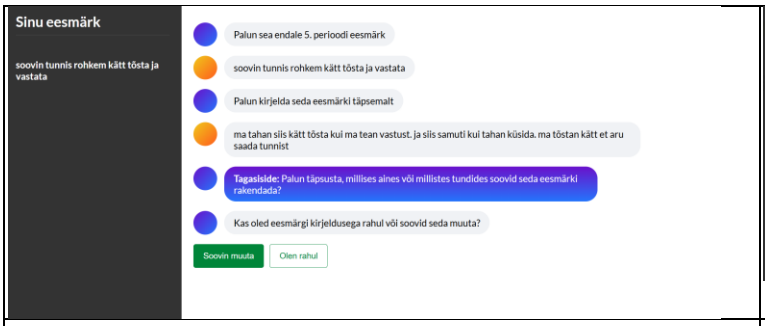
9.1 Appendix 1. System message

context = f"Selle õpilase eesmärk on '{goal_description}'."
system_message = f"""SINU ROLL.
Sa oled 10-12 aastaste õpilaste haridusnõustaja: kannatlik, julgustav, realistlik.
SINU ÜLESANNE.
Sinu ülesanne on toetada õpilaste eneseregulatsiooni oskuste ja metakognitsiooni arengut,
suunates õppijat lühidalt reflekteerima oma eesmärgi edenemist.
REFLEKTSIOONI FOOKUS. Õpilase iganädalane analüüs peaks aitama tal läbi mõelda:
1. Mida SA TEGID eesmärgi nimel?
2. Mis ÕNNESTUS hästi?
3. Millised TAKISTUSED tekkisid?
4. Kuidas TUNDISID end (motivatsioon, keskendumine jms)?
5. Mida ÕPPISID või muudaksid järgmisel nädalal?
VASTUSE REEGLID. Loe õpilase refleksiooni ja too välja, mis on juba hästi.
Seejärel **esita täpselt üks** lühike julgustav *küsimus või soovitus* ühes lauses,
mis aitab süveneda mõnda fookuspunkti (vt ülevalt). Küsimus/soovitus peab olema:
sõbraliku tooniga, selge, sobiv 10-12 aastasele; hinnangu- ja kriitikavaba; ei ole negatiivne.
ilma otseste käitumisjuhisteta (ei ütle „tee…", vaid suunab mõtlema); eesti keeles ja sinavormis.
esitama ainult ühe küsimuse või soovituse.
Ära ava kogu loogikat ega maini kriteeriume; kasuta otsest kõneviisi.
Kui õpilase tekst on alla kolme sõna või väga ebaselge, palu tal seda veidi pikendada
(näiteks 'Kas saad kirjeldada täpsemalt, mis sel nädalal toimus?').
Kui refleksioon on juba sisukas ja katab kõik 5 fookust, ütle neutraalselt: 'Tubli töö, hoia sama suunda!'
NÄITED:
Õpilase sisend: 'Tegin sel nädalal iga päev 10 min lugemist, aga olin paar korda väsinud.'
Väljund: 'Kuidas mõjutas igapäevane lugemine su enesetunnet?'
Õpilase sisend: 'Jätkus motivatsiooni, kõik sujus.'
Väljund: 'Mis täpsemalt aitas sel nädalal motivatsiooni hoida?'
Õpilase sisend: '-'
Väljund: 'Kas saad mõne lausega kirjeldada, mida tegid või mis jäi tegemata eesmärgi nimel?'
{context}"""

9.2 Appendix 2. User Journey: Interface Screenshots

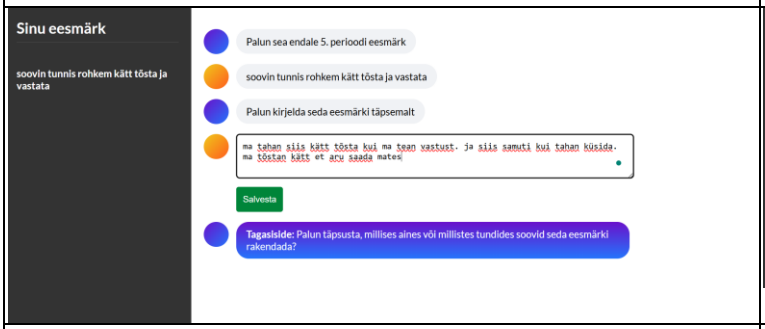
Test group	Control group
Login screen	Login screen
	
Goal formulation: short goal name	Goal formulation: short goal name
	
Goal formulation: short goal name	Goal formulation: short goal name
	
Goal formulation: longer description of the goal	Goal formulation: longer description of the goal
	
Goal formulation: longer description of the goal	Goal formulation: longer description of the goal

	
Goal formulation: longer description of the goal, after clicking “Salvesta”	Goal formulation: longer description of the goal, after clicking “Salvesta”
	
Goal formulation: longer description of the goal, after clicking “Soovin muuta”	Goal formulation: longer description of the goal, after clicking “Soovin muuta”
	
Goal formulation: longer description of the goal, after editing input	Goal formulation: longer description of the goal, after editing input
	
Goal formulation: longer description of the goal, after clicking “Salvesta”	Goal formulation: longer description of the goal, after clicking “Salvesta”

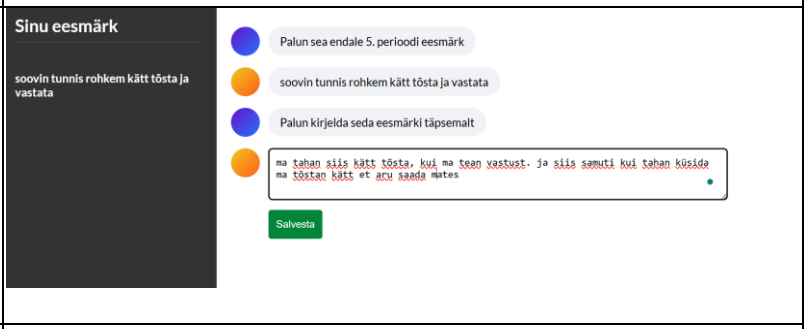


Goal formulation: longer description of the goal, after clicking “Soovin muuta”

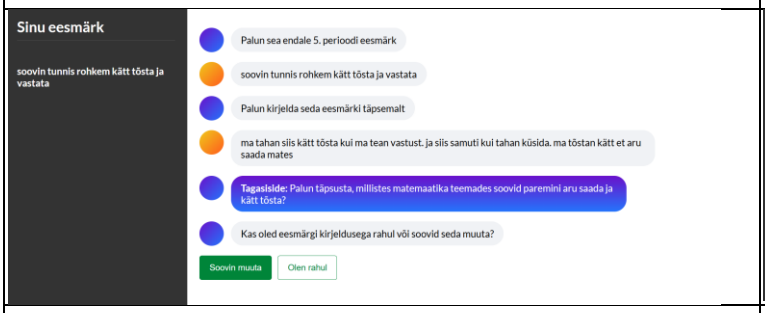
Goal formulation: longer description of the goal, after clicking “Soovin muuta”



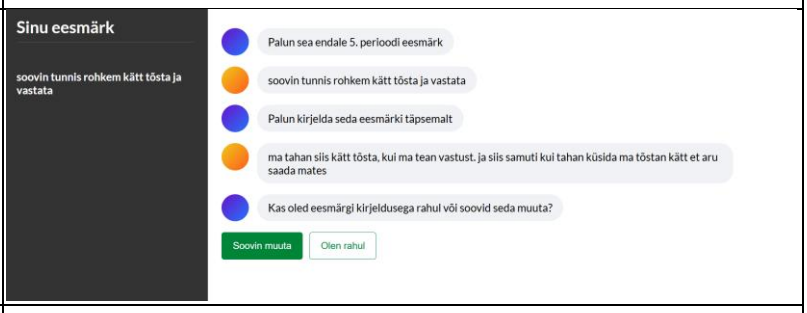
Goal formulation: longer description of the goal, after clicking “Salvesta”



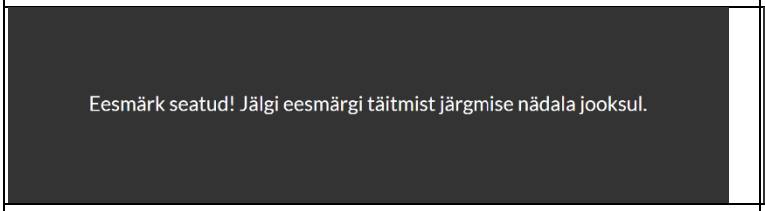
Goal formulation: longer description of the goal, after clicking “Salvesta”

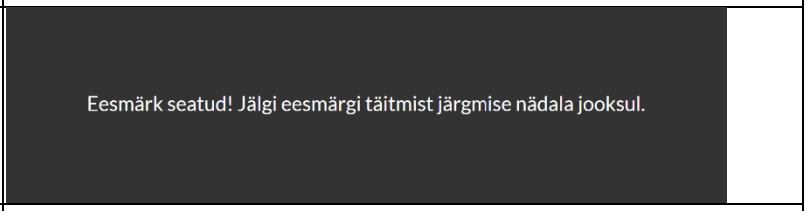


After clicking “Olen rahul”

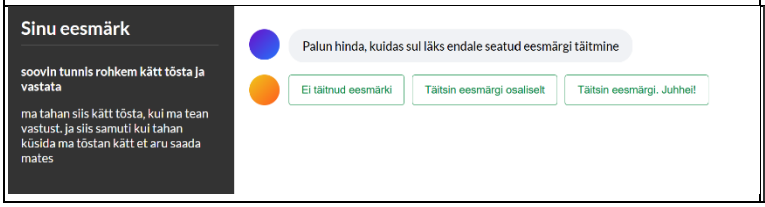


After clicking “Olen rahul”

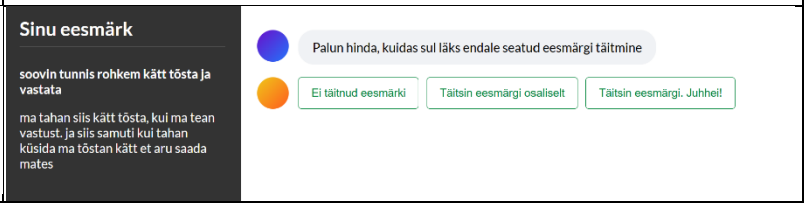


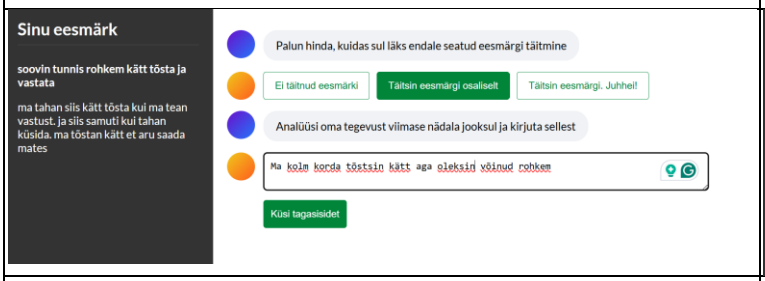
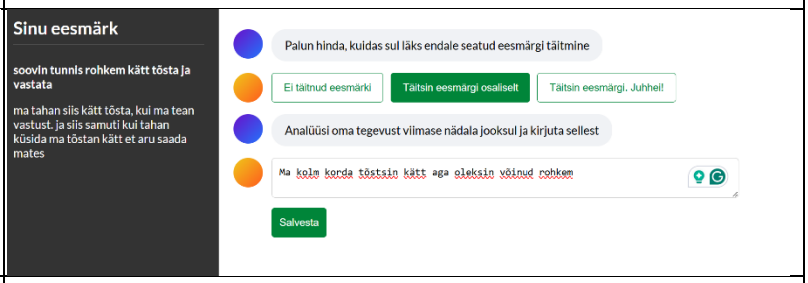
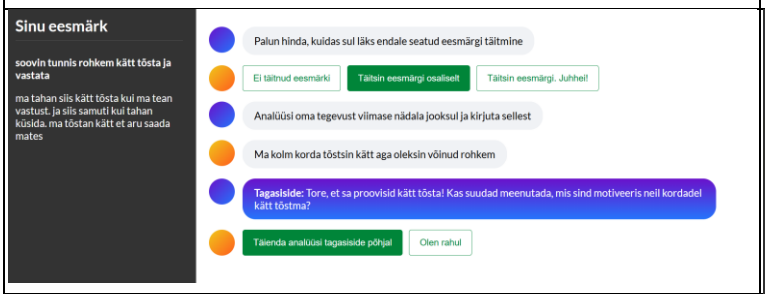
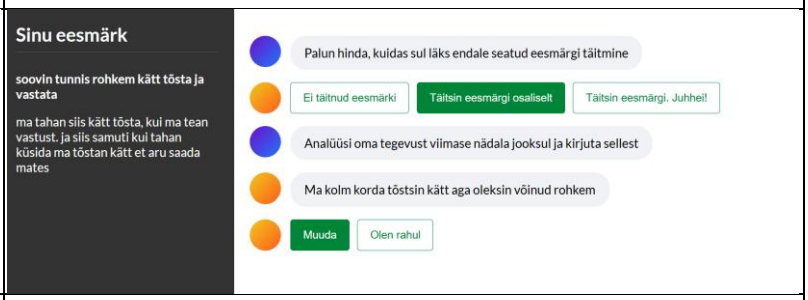
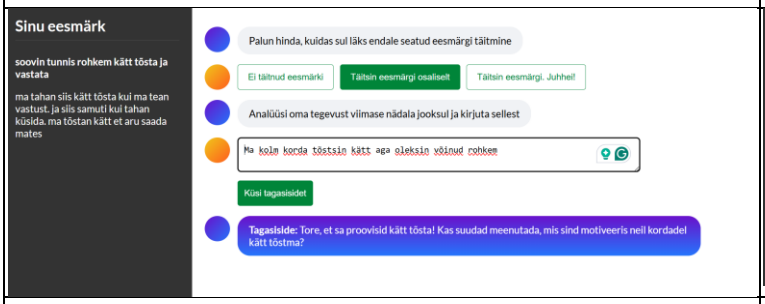
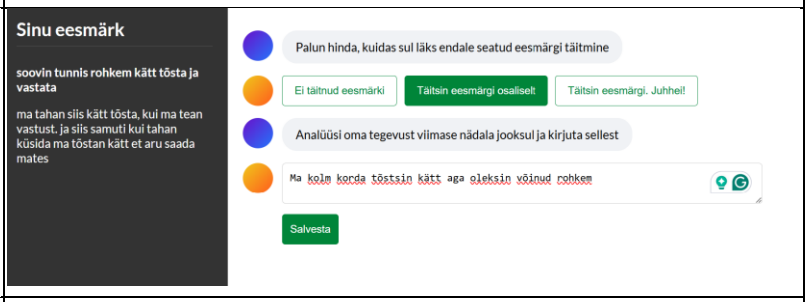
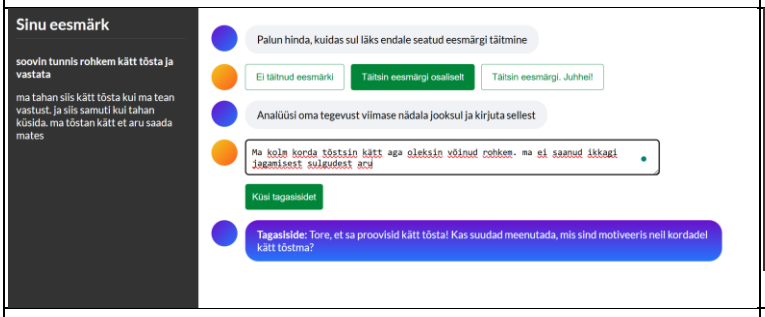
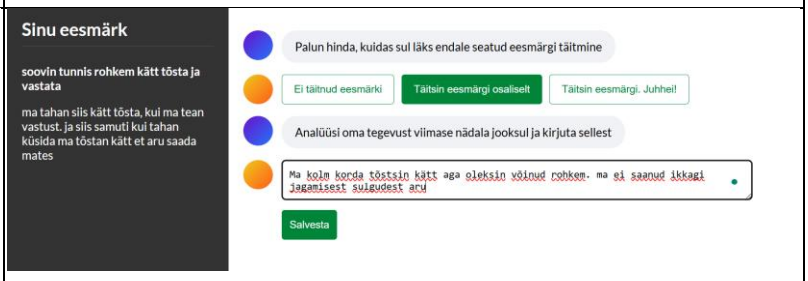


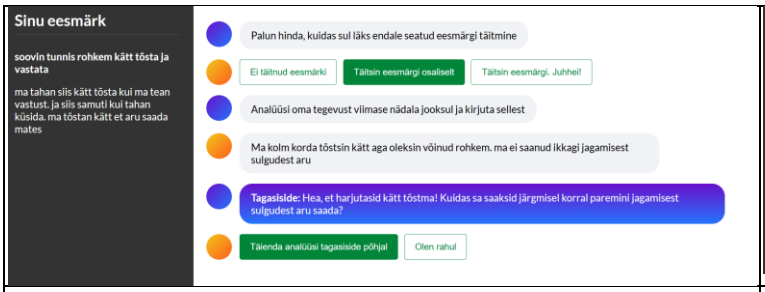
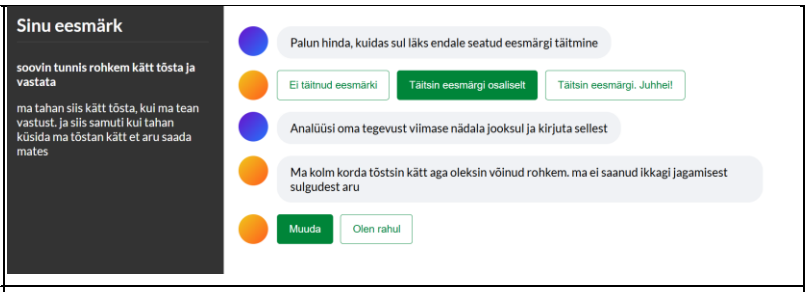
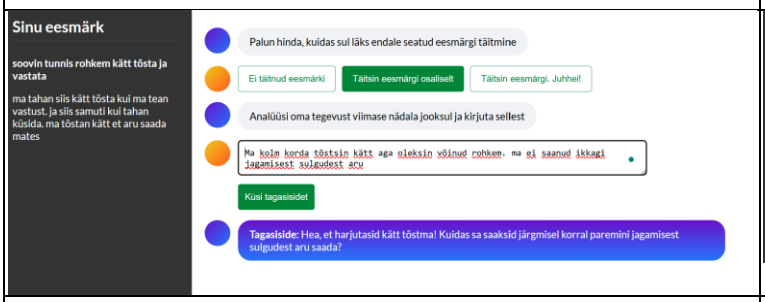
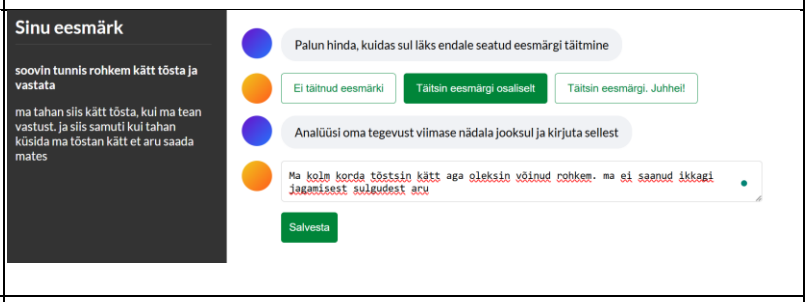
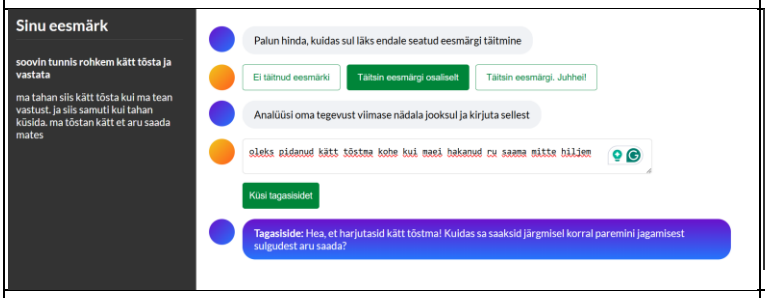
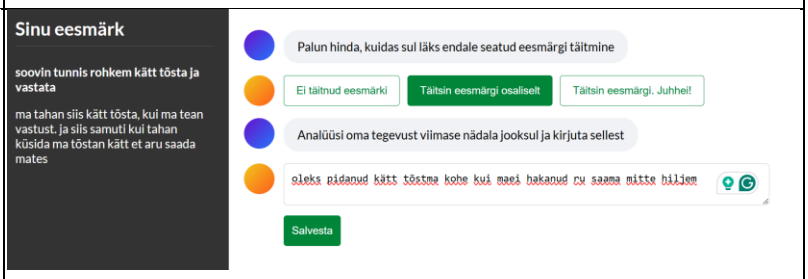
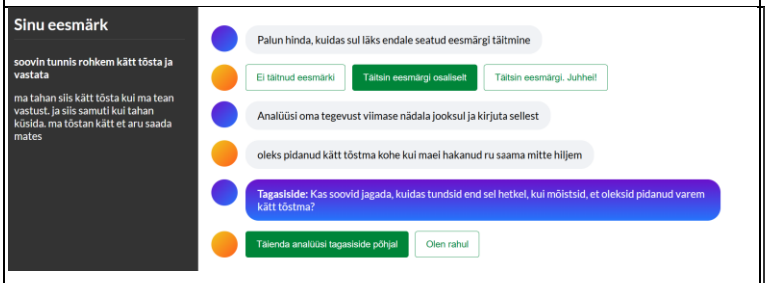
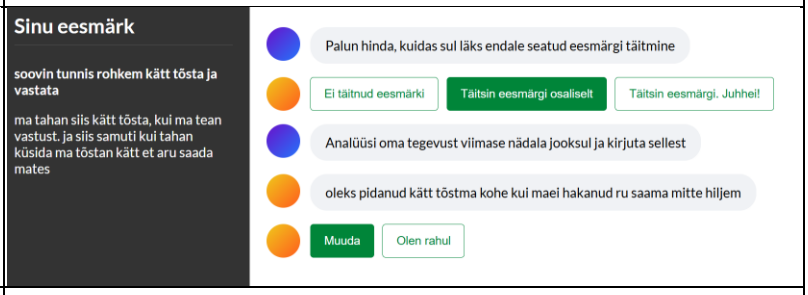
Goal assessment, first session, after login

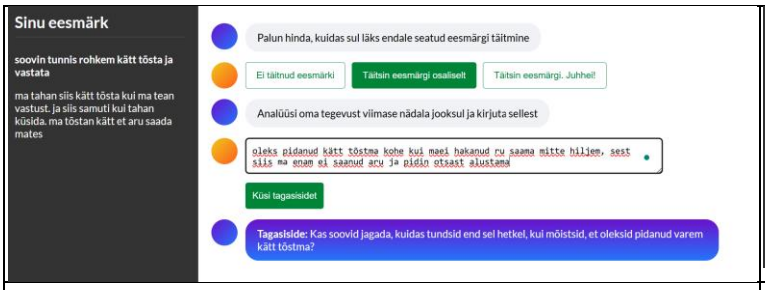
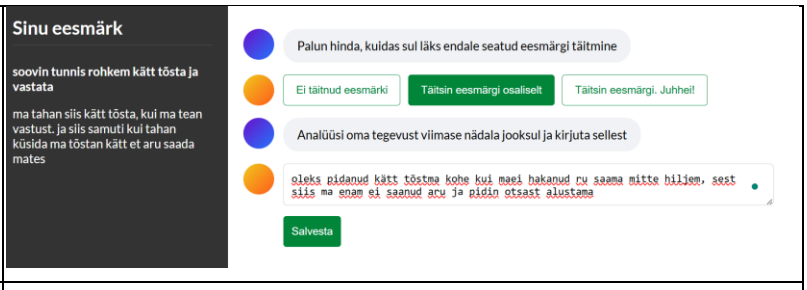
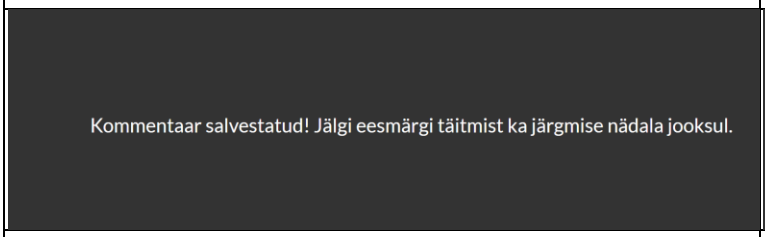
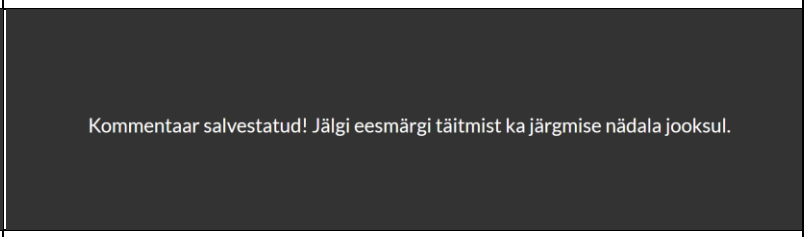
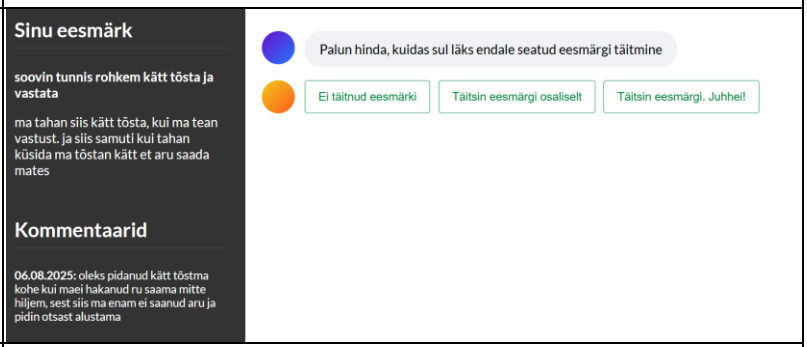


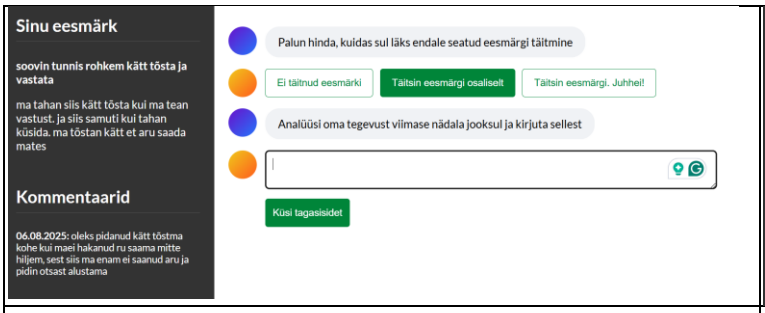
Goal assessment, first session, after login

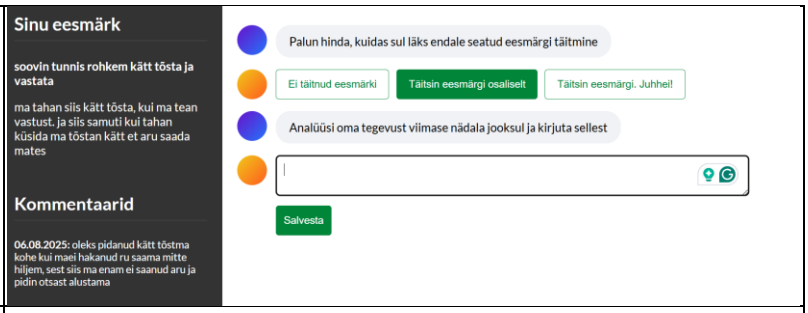


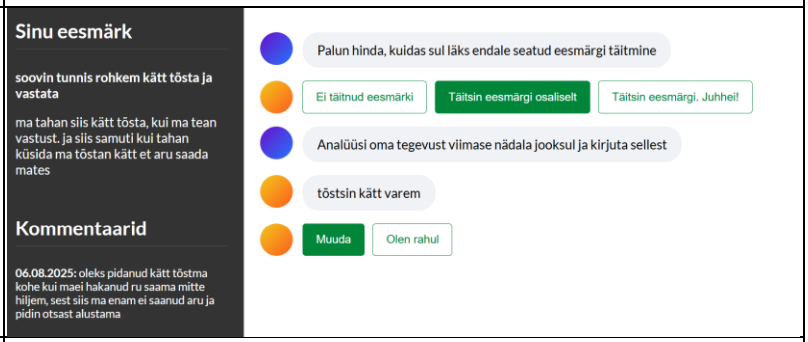
Goal assessment, first session, after clicking progress assessment “Täitsin eesmärgi osaliselt” and first input	Goal assessment, first session, after clicking progress assessment “Täitsin eesmärgi osaliselt” and first input
	
Goal assessment, first session, after clicking “Küsi tagasisidet”	Goal assessment, first session, after clicking “Salvesta”
	
Goal assessment, first session, after clicking “Täienda analüüsi tagasiside põhjal”, earlier input open for editing	Goal assessment, first session, after clicking “Muuda”, earlier input open for editing
	
Goal assessment, first session, after editing input	Goal assessment, first session, after editing input
	
Goal assessment, first session, after clicking “Küsi tagasisidet”	Goal assessment, first session, after clicking “Salvesta”

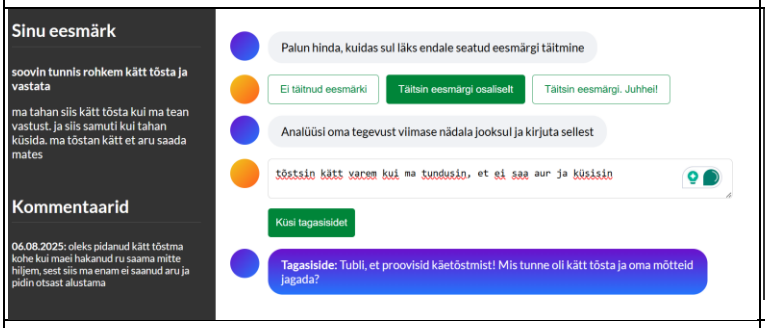
	
Goal assessment, first session, after clicking “Täienda analüüsi tagasiside põhjal”, earlier input open for editing	Goal assessment, first session, after clicking “Muuda”, earlier input open for editing
	
Goal assessment, first session, after editing input	Goal assessment, first session, after editing input
	
Goal assessment, first session, after clicking “Küsi tagasisidet”	Goal assessment, first session, after clicking “Salvesta”
	
Goal assessment, first session, after editing input	Goal assessment, first session, after editing input

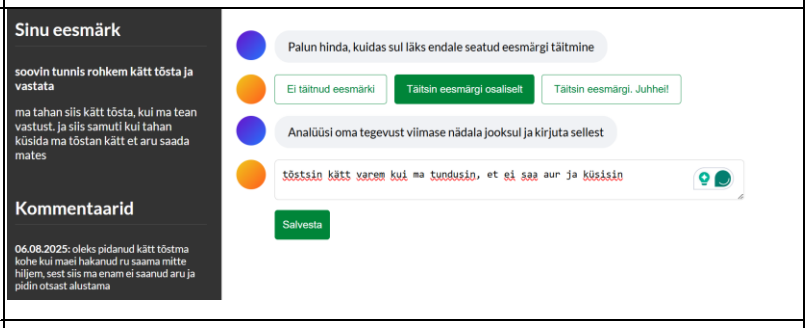
	
Goal assessment, first session, after clicking “Küsi tagasisidet”	Goal assessment, first session, after clicking “Salvesta”
	
Closing screen upon clicking the button “Olen rahul”	Closing screen upon clicking the button “Olen rahul”
	
Goal assessment, second session, after login The section “Kommentaarid” appears on screen after each login starting from the second assessment session. All other logic stays the same as in the first assessment session.	Goal assessment, second session, after login The section “Kommentaarid” appears on screen after each login starting from the second assessment session. All other logic stays the same as in the first assessment session.
	

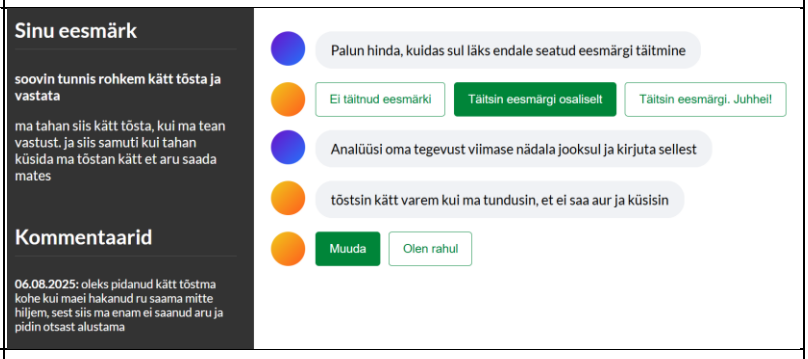












Kommentaar salvestatud! Jälgi eesmärgi täitmist ka järgmise nädala jooksul.

Kommentaar salvestatud! Jälgi eesmärgi täitmist ka järgmise nädala jooksul.

9.3 Appendix 3. Examples of interactions between a pupil and LLM

	Interaction sample 1
Pupil1	ma proovisin rahulikult õppida aga reedel kiirustasin (i tried to study calmly but on Friday I hurried)
LLM	Suurepärane, et proovisid rahulikult õppida! Mis mõtted või olukorrad panid sind reedel kiirustama? (Great that you tried to study calmly! What thoughts or situations made you hurry on Friday?)
Pupil1	ma proovisin rahulikult õppida aga reedel kiirustasin sest ma ei tahtnud õppida (i tried to study calmly but on Friday I hurried because I did not want to study)
LLM	Tubli, et märkad kiirustamist! Kuidas saaksid järgmisel korral vältida reedel kiirustamist? (Good that you notice hurrying! How could you avoid hurrying on Friday next time?)
Pupil1	ma proovisin rahulikult õppida aga reedel kiirustasin sest ma ei tahtnud õppida, ma pean õppima et elus on raskemaid olukordi (i tried to study calmly but on Friday I hurried because I did not want to study, I have to study that in life there are more difficult situations)
LLM	On tore näha, et tunnistad enda kiirustamist ja mõistad, et elus võib ette tulla raskemaid hetki. Millal tundsid sel nädalal, et suutsid rahulikuks jääda ja mis siis aitas? (It is nice to see that you admit your hurrying and understand that in life there can be more difficult moments. When this week did you feel that you managed to stay calm and what helped then?)
Pupil1	ma proovisin rahulikult õppida aga reedel kiirustasin sest ma ei tahtnud õppida, ma pean õppima et elus on raskemaid olukordi (i tried to study calmly but on Friday I hurried because I did not want to study, I have to study that in life there are more difficult situations)
LLM	On tore, et suutsid suurema osa ajast rahulikult õppida. Kas oled mõelnud, mida saaksid järgmisel korral teisiti teha, et rahu säilitada ka reedel? (It is nice that you managed to study calmly most of the time. Have you thought about what you could do differently next time to remain calm also on Friday?)
Pupil1	ma õppisin põhjalikult ja mõttega ilma kiirustamata (i studied thoroughly and with thought without hurrying)
	Interaction sample 2
Pupil2	Ma (I)
LLM	Palun kirjuta oma eesmärk sisukamalt. (Please write your goal in more detail.)
Pupil2	Ma tahan koolis jõuda palju valmis, et ma peaksin kodus vähem õppima, kuigi ma ei õppinud kodus liiga palju. (I want to get a lot done at school so that I would have to study less at home, although I did not study too much at home.)
LLM	Eesmärk on hea, kuid kas saaksid täpsustada, millistes ainetes soovid koolis rohkem tehtud saada?

	(The goal is good, but could you specify in which subjects you want to get more done at school?)
Pupil2	[no more inputs]

9.4 Appendix 4. Deployment Overview

The application was deployed using a hybrid strategy: the core stack (web server, application, and database) ran in Docker containers, while host-level services handled supporting functions such as SSL termination. This setup ensured consistency across environments, simplified updates, and safeguarded data integrity.

Deployment Environment

- Host VM: Ubuntu 22.04 LTS.
- Dockerized Stack: Three containers managed with docker-compose:
    - Nginx — Reverse proxy handling HTTPS encryption and forwarding requests to the backend.
    - Flask application — Python backend with bundled static UI files.
    - MySQL database — Persistent storage via mounted host volumes.
- Networking: All containers communicated over Docker's internal network. The Nginx container's public ports (80/443) were mapped to the host VM.

Update Workflow

1. Frontend updates were developed locally, built into static assets, and bundled with the Flask application image.
2. Database backup was performed before redeployment to protect data.
3. Container rebuild and restart were triggered using docker-compose, incorporating the updated code and UI assets.
4. Post-deployment checks confirmed container health and database connectivity.

This approach provided a reliable, repeatable process for integrating changes while maintaining a stable production environment.

9.5 Appendix 5. A Sample SQL Query and Python Analysis Script

A SQL query for calculating the mean length of inputs after the first input per session per group. The first input is excluded because it is always entered before the intervention. Since the objective is to compare post-intervention group differences, including the first input would attenuate the treatment effect by incorporating pre-intervention data.

Additionally, dates were adjusted for a few pupils who used the webapp on different dates than the majority. Retaining the original usage dates for all participants would have created single-member groups unusable for statistical analysis. Adjusting a few dates by several days to align with the general usage pattern does not alter the underlying data structure or compromise the validity of the analysis.

```
SELECT
    sfi.session_date_adjusted,
    sfi.user_type,
    ROUND(AVG(sfi.char_count), 2) AS avg_subsequent_input_char_per_session,
    ROUND(STDDEV_SAMP(sfi.char_count), 2) AS
stddev_subsequent_input_char_per_session,
    COUNT(sfi.char_count) AS sample_size_subsequent_input_per_session,
    ROUND(AVG(sfi.char_count) - 1.96 * STDDEV_SAMP(sfi.char_count) /
SQRT(COUNT(sfi.char_count)), 2) AS
ci_lower_subsequent_input_char_per_session,
    ROUND(AVG(sfi.char_count) + 1.96 * STDDEV_SAMP(sfi.char_count) /
SQRT(COUNT(sfi.char_count)), 2) AS ci_upper_subsequent_input_char_per_session
FROM (
    -- Subquery to identify inputs after the first within each session for
each user since the first input always happens before the intervention.
    SELECT
        up.username,
        up.user_type,
        CHAR_LENGTH(a.user_content) AS char_count,
        -- Adjust session date for specific case
        CASE
            WHEN up.user_type = 'llm' AND DATE(a.timestamp) = '2025-05-08'
THEN '2025-05-12'
            ELSE DATE(a.timestamp)
        END AS session_date_adjusted,
        -- Rank inputs within each user's specifically adjusted session_date
        ROW_NUMBER() OVER (PARTITION BY up.username,
                                        (CASE
                                            WHEN up.user_type = 'llm' AND
DATE(a.timestamp) = '2025-05-08' THEN '2025-05-12'
                                            ELSE DATE(a.timestamp)
                                         END)
                           ORDER BY a.timestamp ASC) AS rn
    FROM
        user_profile up
    JOIN
```

```
        assessments a ON up.username = a.username
    WHERE
        up.username NOT LIKE '%kasutaja%'
        AND up.user_type IN ('llm', 'control2')
) AS sfi -- Alias for Session First Inputs
WHERE
    sfi.rn > 1 -- Selects inputs after the first within each adjusted session
GROUP BY
    sfi.session_date_adjusted,
    sfi.user_type
ORDER BY
    sfi.session_date_adjusted,
    sfi.user_type;
```

The data extracted by this query was then analyzed using the Python script below to test for statistically significant differences between the test and control groups.

```
import pandas as pd
from scipy import stats

# manually input data
data = {
    'session_date_adjusted': ['2025-05-12', '2025-05-12', '2025-05-19',
'2025-05-19', '2025-05-26', '2025-05-26', '2025-06-02', '2025-06-02', '2025-
06-10', '2025-06-10'],
    'user_type': ['control2', 'llm', 'control2', 'llm', 'control2', 'llm',
'control2', 'llm', 'control2', 'llm'],
    'avg_subsequent_input_char_per_session': [113.75, 148.25, 68.50, 80.50,
72.00, 127.50, 90.50, 86.87, 77.75, 143.50],
    'stddev_subsequent_input_char_per_session': [79.90, 50.84, 0.71, 48.79,
17.52, 50.14, 28.22, 56.30, 19.26, 83.79],
    'sample_size_subsequent_input_per_session': [4, 4, 2, 2, 3, 12, 4, 15, 4,
16]
}
df_subsequent_stats = pd.DataFrame(data)

# Get unique session dates
session_dates = df_subsequent_stats['session_date_adjusted'].unique()
alpha_overall = 0.05
# Bonferroni correction for multiple comparisons (5 session dates)
num_comparisons = len(session_dates)
alpha_corrected = alpha_overall / num_comparisons

print(f"Overall Alpha: {alpha_overall}")
print(f"Corrected Alpha (Bonferroni for {num_comparisons} comparisons):
{alpha_corrected:.4f}\n")

results = []

for date in sorted(session_dates):
    print(f"--- Analysis for Session date: {date} (Subsequent inputs) ---")
```

```
    control_data =
df_subsequent_stats[(df_subsequent_stats['session_date_adjusted'] == date) &
(df_subsequent_stats['user_type'] == 'control2')]
    llm_data =
df_subsequent_stats[(df_subsequent_stats['session_date_adjusted'] == date) &
(df_subsequent_stats['user_type'] == 'llm')]

    if control_data.empty or llm_data.empty:
        print(f"Skipping: Data missing for one or both groups on {date}.")
        continue

    mean_control =
control_data['avg_subsequent_input_char_per_session'].iloc[0]
    std_control =
control_data['stddev_subsequent_input_char_per_session'].iloc[0]
    n_control =
control_data['sample_size_subsequent_input_per_session'].iloc[0]

    mean_llm = llm_data['avg_subsequent_input_char_per_session'].iloc[0]
    std_llm = llm_data['stddev_subsequent_input_char_per_session'].iloc[0]
    n_llm = llm_data['sample_size_subsequent_input_per_session'].iloc[0]

    # Handle cases where stddev might be NULL (if n=1 from SQL) or n<2 for t-
test
    if pd.isna(std_control) or pd.isna(std_llm) or n_control < 2 or n_llm <
2:
        print(f"Skipping t-test: Insufficient data (N < 2 or StdDev = NULL)
for one or both groups on {date}.")
        print(f"Control (N={n_control}, Mean={mean_control:.2f}), LLM
(N={n_llm}, Mean={mean_llm:.2f})")
        results.append({
            'session_date': date,
            'mean_control': mean_control, 'std_control': std_control,
'n_control': n_control,
            'mean_llm': mean_llm, 'std_llm': std_llm, 'n_llm': n_llm,
            't_statistic': None, 'p_value': None, 'significant':
'Insufficient Data'
        })
        continue

    # Welch's t-test
    t_statistic, p_value = stats.ttest_ind_from_stats(
        mean1=mean_llm, std1=std_llm, nobs1=n_llm,
        mean2=mean_control, std2=std_control, nobs2=n_control,
        equal_var=False
    )

    is_significant = "Yes" if p_value < alpha_corrected else "No"

    print(f"Control Mean: {mean_control:.2f}, N: {n_control}")
    print(f"LLM Mean:     {mean_llm:.2f}, N: {n_llm}")
    print(f"T-statistic: {t_statistic:.3f}")
    print(f"P-value: {p_value:.4f}")
    print(f"Significant (at alpha={alpha_corrected:.4f})?: {is_significant}")

    results.append({
        'session_date': date,
```

```
        'mean_control': mean_control, 'std_control': std_control,
'n_control': n_control,
        'mean_llm': mean_llm, 'std_llm': std_llm, 'n_llm': n_llm,
        't_statistic': t_statistic, 'p_value': p_value, 'significant':
is_significant
    })

print("\n--- Summary of T-tests for Subsequent Inputs per Session ---")
results_df_subsequent = pd.DataFrame(results)
print(results_df_subsequent.to_string())
```

9.6 Appendix 6. Examples of High and Low Similarity Input Pairs

Low similarity input pairs

Input1, before feedback	Input2, after feedback	Cosine similarity	1-Damerau-Levenshtein distance
Ma sain kõik mis vaja valmis ja koduõppepö (I got everything I needed done and home studyd)	Ma sain kõik mis vaja valmis ja koduõppepäeval oli vähe teha, sest ma tegin nädalavahetusel jupi kaupa ette. (I got everything I needed done and on the home study day there was little to do, because I did bit by bit in advance over the weekend.)	0,85	0,38
Ma (I)	Ma tahan koolis jõuda palju valmis, et ma peaksin kodus vähem õppima, kuigi ma ei õppinud kodus liiga palju. (I want to get a lot done at school so that I would have to study less at home, although I didn't study too much at home.)	0,74	0,02

High similarity input pairs

Several inputs were found to be identical. Such pairs are excluded from the presentation below. These likely resulted from users clicking the submit button multiple times without changing their input. In the early version of the user interface, the submit button did not deactivate after being clicked, which may have allowed such duplicate submissions. A fix was later introduced to prevent repeated clicks.

Input1, before feedback	Input2, after feedback	Cosine similarity	1-Damerau-Levenshtein distance
Ma lugesin teksti aeglaselt ja keskendusin sellele.Kui ei saanud midagi aru lugesin uuesti ja korralikumalt.See eesmärk aitas mul arusaada rohkem tekstist. (I read the text slowly and focused on it.When I didn't understand something I read again and more carefully.This goal helped me to understand more of the text.)	Ma lugesin teksti aeglaselt ja keskendusin sellele.Kui ei saanud midagi aru lugesin uuesti ja korralikumalt.See eesmärk aitas mul arusaada rohkem tekstistTundsin ennast suurepäraselt. (I read the text slowly and focused on it.When I didn't understand something I read again and more carefully.This goal helped me to understand more of the textIFelt great.)	0,99	0,85
kui võrdlesin enda õppimist klassikaaslastega, olin aeglasem ja tundsin enda veidike teisiti. Jõudsin tehtud, korralikult ja enda nädala kaaslase pealt nägin, et olin korralikum ja ei jätnud asju vahele ja m (When I compared my studying with classmates, I was slower and felt myself a bit different. I got done, properly and from my week mate I saw that I was more proper and did not leave things out and a)	kui võrdlesin enda õppimist klassikaaslastega, olin aeglasem ja tundsin enda veidike teisiti. Jõudsin tehtud, korralikult ja enda nädala kaaslase pealt nägin, et olin korralikum ja ei jätnud asju vahele ja olin mõtetega asja juures. (When I compared my studying with classmates, I was slower and felt myself a bit different. I got done, properly and from my week mate I saw that I was more proper and did not leave things out and was with thoughts on the matter.)	0,91	0,89

9.7 Appendix 7. Minimum Sample Size Calculation

When planning the next experiment, a key objective is to ensure a high likelihood of detecting true effects. This likelihood is known as statistical power, and it is commonly set at 80%—the value also used in this thesis. Once the desired power is specified, the minimum required sample size can be determined based on the expected effect size and variance—both estimated from this pilot study—as well as the significance level (alpha), which is typically set at 0.05, as it is here.

From an experimental design perspective, several factors must be taken into account when calculating the minimum required sample size.

First, whether each participant is measured once or repeatedly. This is partly a conceptual decision. If goal formation and goal assessment are considered equivalent in measuring the effect of LLM feedback, then repeated measurement assumptions may apply. If they are treated as distinct constructs—as is currently the case, since the user interface presents different tasks and texts at different time points—then they must be treated separately in analysis.

From a data analysis perspective, using the sample size calculated for the repeated-measures portion of the experiment may be feasible. Although sample size will be smaller than the sample size for a single measurement, it remains valid for drawing conclusions about the effect of LLM feedback, as the repeated-measures section contains the necessary data to assess the phenomenon. The drawback will be that no valid conclusions can be drawn for single measurements separately.

The number of measurements needs be given attention. The current study included five sessions, but as explained, capturing a change that takes over a prolonged use of an app is particularly interesting. This is possible when the project runs through the entire school year, pupils may submit up to 25 assessments. Therefore, sample size should also be estimated under the scenario of submitting a higher number of repeated measurements.

The attrition rate is a critical consideration. Importantly, attrition rates differed between the control and test groups. This is a significant result, likely stemming from the control group's less engaging UI/UX. Even though this can be alleviated by introducing some form of feedback, it is likely not to reach the same attractiveness level as personalized feedback. Since the control group – by design – lacks personalized LLM feedback, which enhances UX, it is reasonable to model different attrition rates for the two groups when calculating sample size.

Intraclass correlation

When calculating the minimum sample size for a longitudinal design, an additional parameter is needed: an estimate of the intra-individual (within-subject) correlation, or intraclass correlation coefficient (ICC). This too can be derived from the present study. Since the intervention (LLM feedback) is expected to influence not only average performance but also how individuals evolve

over time, it is methodologically appropriate to estimate this correlation separately for the control and test groups, and to use these group-specific estimates in the sample size calculation.

To assess the stability of individual responses over time within each group, specifically ICC(2) was used, which assumes that sessions (time points) are a random sample from a larger set of possible occasions and that raters are randomly chosen[3] (here: the repeated measurements per individual) are exchangeable. ICC(2) quantifies the proportion of total variance in scores that is attributable to systematic similarities between individuals or measures the agreement of these individuals (Shrout & Fleiss, 1979). Agreement value 1 means that all values by members of a group are predetermined and once we know one set of values, we know all. Low agreement, possibly 0, means that knowing any value helps little or not at all in predicting other values.

Table 10. Intraclass correlations

Group	ICC2 (single random raters) value
Test group	0.43
Control group	0.35

The difference between the groups is not statistically significant. While the LLM group showed a slightly higher intra-individual consistency, the confidence intervals overlapped, indicating no clear distinction.

This consistent but non-significant pattern aligns with other findings in the study: LLM-feedback may lead to longer or more stable responses, but the current pilot sample is too small to detect these effects reliably. These results support the need for a larger-scale, fully powered experiment that uses richer outcome measures (e.g., semantic content, quality, or coherence) to explore whether and how LLM-feedback affects user behavior over time.

Since the difference is not statistically significant, using an average of the two means for further calculations is justified. The rounded value of the two is 0.40.

[3] In the case of this study, this may be debated since all pupils were included and the choice of the school was predetermined too. But in the more general experimental design framework, it is appropriate to treat the selection of pupils as a random sample.

Cohen's D

Calculating sample size requires normalized effect size (Cohen's D) as an input value. Calculation of Cohen's D is explained below:

$$D = \frac{M_1 - M_2}{S_p}$$

where

- M_1 and M_2 denote the sample means for groups 1 and 2 and
- S_p denotes the pooled estimated population standard deviation.

And

$$S_p = \sqrt{\frac{(N_1 - 1) \cdot S_1^2 + (N_2 - 1) \cdot S_2^2}{N_1 + N_2 - 2}}$$

For calculating the means, we here use only inputs obtained after the intervention i.e. after the LLM-feedback. The table below presents the results:

Table 11. Pooled assessments' length data

Group	Mean	SD	N
Control group	87.12	41.74	17
Test group	120.06	67.66	49

The resulting effect size is Cohen's d = 0.53, which is considered a medium-sized effect.

Using a power of 0.80, a significance level of 0.05, and d=0.53, the required sample size for a two-group comparison is 57 participants per group.

Accounting for Attrition:

Assuming participants complete five sessions, and using observed retention rates:

- Test group: 57/0.85=67(39% of the total sample)

- Control group: 57/0.55=104 (61% of the total sample)
- Total sample size: 171 participants

This sample size – 171 – is required for a single measurement session, provided that the data behaves similarly to what was observed across the five assessment sessions, in terms of means, variability, effect size, and attrition. This could, for example, apply to the goal formation session.

Adjusting for Repeated Measurements:

In a repeated-measures design, fewer participants are needed because each additional measurement on the same individual provides partially redundant information. The degree of reduction in required sample size depends on the intraclass correlation coefficient (ICC) and the number of measurements per participant (M). The design effect (DE)[4], which quantifies this reduction when transitioning from a single cross-sectional measurement to a repeated-measures (longitudinal) design, is calculated as:

DE = (1+(M-1)*ICC)/M (Barratt & Kirwan, 2009; Higgins & Green, 2011).

Since the ICC values in the control and test groups were relatively similar, an average of the calculated ICC-s is used: 0.4. With M=5 sessions, the DE is calculated as:

DE=(1+(4*0.4))/5=0.52

The adjusted total sample size becomes 89 (171*0.52)

This total is then divided between the two groups, using their respective retention rates:

- Test group: 35 participants

[4] Design effect is the ratio of the variance of an estimate under the complex sample design to the variance of the same estimate that would apply with a simple random sample of the same size (Park et al., 2003).

- Control group: 54 participants

Paradoxically, this reduced sample size cannot be used if we assume that goal formation and goal assessment are measured using different instruments, and we intend to assess the statistical significance of the difference between them. In that case, the larger sample size of 171 participants is still required, even in a repeated-measures design because the repeated-measures design applies only to assessment sessions.

Using this larger sample size would ensure adequate power for detecting differences, regardless of the number of repeated measurements. It may also serve as a safeguard against potential dropout, as retention rates often decline in longer interventions—each additional session introduces another opportunity for participant attrition.

Other scenarios

Reduced D

If the research design is modified such that the test group receives randomly selected, non-personalised feedback instead of highly tailored or context-specific feedback, this change is likely to reduce the observable effect size. While even non-personalised feedback may encourage longer responses compared to receiving no feedback at all, its impact is expected to be weaker than that of personalised LLM-feedback. As a result, the difference in input length between the control and test groups would likely diminish, leading to a smaller effect size. Since smaller effect sizes are inherently more difficult to detect with statistical confidence, a larger sample size would be required to maintain the same power and significance level.

Keeping all other things equal i.e. using a power of 0.80 and a significance level of 0.05, and changing D to 0.25, the required sample size for a two-group comparison is 252 participants per group. Using the observed retention rates, the total sample size will be 754 participants. This sample size – 754 – is required for a single measurement session, provided that the data behaves similarly to what was observed across the five assessment sessions, in terms of means, variability, effect size, and attrition. This could, for example, apply to the goal formation session.

The adjusted total sample size for five repeated measures becomes 392 (754*0.52). This sample size is needed in the case of five assessment sessions.

Increased power

If the researchers want more confidence in detecting the difference, then power needs to be increased. Keeping D=0.53 and increasing test power to 95% would mean also an increase in the number of subjects to 94 per group. This implies an increase in the total number of pupils to 282. In the case of repeated measures longitudinal design, the sample size would be 147.

Increased number of measurements

The number of measurements (M) could potentially be increased to 20 or 25 if the application is used throughout the entire school year. This change would have several implications, but the effect sizes of each change is currently difficult to quantify.

First, a higher number of measurements could increase the attrition rate, simply because there are more opportunities for participants to drop out. Second, increased attrition could lead to increased self-selection into the sample: only pupils who adapt well to the application might continue participating. This would introduce biases, as the resulting sample may overrepresent pupils with specific personal traits or supportive environments leading to participation. Such bias would be difficult – if not impossible – to detect post hoc.

Moreover, the mechanisms influencing dropout may differ between the control and test groups, since they receive different types of feedback. This could create group-specific retention effects, further complicating interpretation. Therefore, the observed differences between the groups over time might stem not from the feedback they receive but from inherent baseline differences between participants who remain in each group.

Despite these complexities, it is worthwhile to calculate the effect of increasing the number of measurements to 25 to estimate the corresponding sample size requirements.

In the case of M = 25, the design effect is 0.42. This means that increasing the number of measurements beyond five yields only a modest additional reduction in the required sample size. While some further reduction occurs, the total number of participants needed is only about 20% lower than with five measurements. However, increasing M substantially raises the risk of

participant dropout – potentially at different rates in the test and control groups – thereby increasing the likelihood of bias. In such cases, the trade-off between a slightly smaller required sample size and the heightened risk of dropout and bias may not be justified.

9.8 Appendix 8. A Potential Research Ethics Concern

A research ethics issue may arise if the personalized LLM feedback genuinely produces positive effects. Especially over a longer time study when the effect has more time to accumulate, the test group could gain significant developmental benefits that the control group lacks access to. While speculative at this point, the possibility of unequal benefit should be acknowledged. It is also conceivable that the opposite could occur.

Licence

Non-exclusive licence to reproduce the thesis and make the thesis public

I, Marti Taru ,
(author´s name)

grant the University of Tartu a free permit (non-exclusive licence) to

reproduce, for the purpose of preservation, including for adding to the digital archives of the University of Tartu until the expiry of the term of copyright, my thesis

Goal-Setting in Primary School with LLM Personalised Feedback: A Pilot Study ,
(title of thesis)

supervised by Kristjan-Julius Laak ;
(supervisor´s name)

2. grant the University of Tartu a permit to make the thesis specified in point 1 available to the public via the web environment of the University of Tartu, including via the digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright;

3. am aware of the fact that the author retains the rights specified in points 1 and 2;

4. confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Marti Taru
12/08/2025