PromptMirror: Visualizing LLM Use to Support STEM Student Reflection

Ka Hei Carrie Lau*
carrie.lau@tum.de
Technical University of Munich
Munich Center for Machine Learning (MCML)
Munich, Germany

Nađa Terzimehić*
nadja.terzimehic@tum.de
Technical University of Munich
Munich Center for Machine Learning (MCML)
Munich, Germany

Enkelejda Kasneci
enkelejda.kasneci@tum.de
Technical University of Munich
Munich Center for Machine Learning (MCML)
Munich, Germany

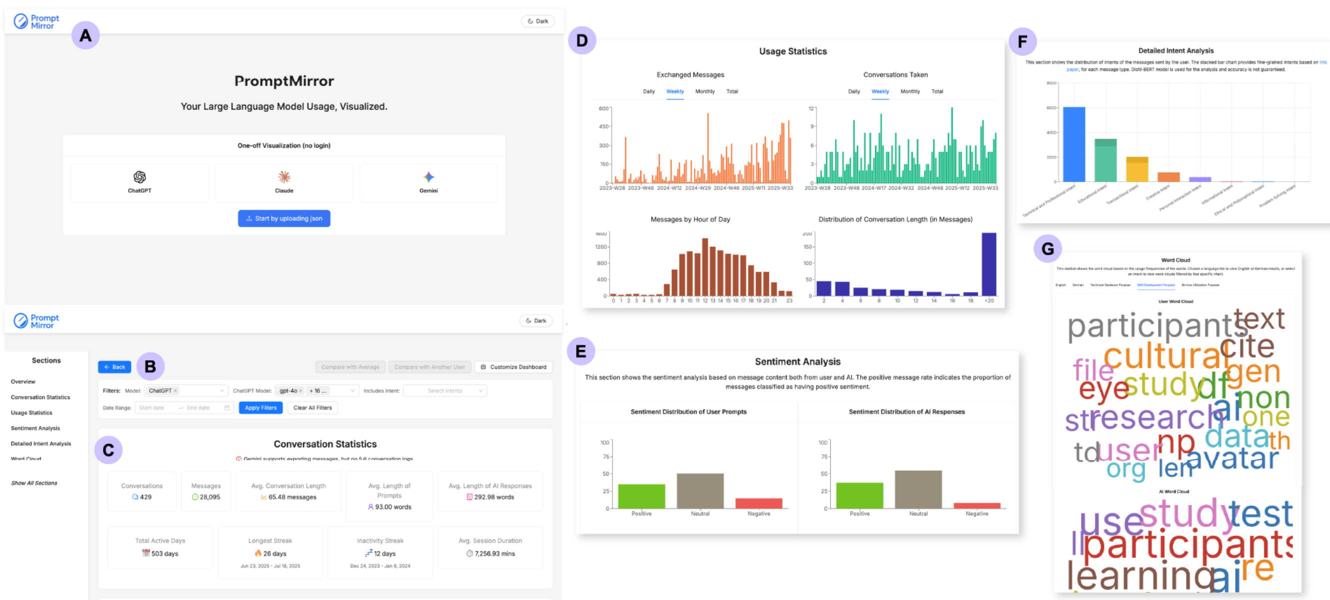


Figure 1: The *PromptMirror* visualizes users' LLM usage through analytics views, while supporting multiple LLM platforms. (A) A one-off, no-login upload page that accepts exports from multiple providers (e.g., ChatGPT, Claude, Gemini). The dashboard then provides five visualization components: (B) a control bar with filters (e.g., model, date range, intent); (C) Conversation Statistics summarizing engagement (e.g., conversations/messages, prompt and response length, active days, streaks, session duration); (D) Usage Statistics showing temporal patterns (e.g., daily/weekly/monthly trends); (E) Sentiment Analysis comparing sentiment in user prompts and AI responses; (F) Detailed Intent Analysis classifying message purposes (e.g., technical, educational, creative); and (G) a Word Cloud view with language selection and intent-specific word clouds. The previous version of *PromptMirror* is shown in Appendix Figure 8.

Abstract

Large language models (LLMs) are increasingly embedded in students' academic work, yet the increasing reliance can undermine learning depth and raise integrity concerns. While reflection has long been studied in HCI to foster awareness and behavior change, little is known about how to support students in reflecting on everyday LLM use. We present *PromptMirror*, a student-facing dashboard that processes LLM conversation logs and visualizes four perspectives, temporal, sentiment, intent, and thematic, to encourage reflection. We informed the design of PromptMirror with two focus groups (one expert and one student with four participants each) and subsequently conducted an online think-aloud with 20 university students who uploaded their own LLM use data. Findings provide preliminary evidence that PromptMirror may support students in recognizing their LLM use estimation gap and engaging in deeper reflection on LLM reliance. Our contributions are twofold: (1) a student-centric reflection system; (2) empirical insights into reflective analytics for everyday LLM tools.

*Both authors contributed equally to this research.

DIS '26, Singapore, Singapore

ACM ISBN 979-8-4007-2563-0/2026/06
https://doi.org/10.1145/3800645.3812863

CCS Concepts

• **Human-centered computing → Visual analytics**; **Empirical studies in HCI**; **User studies**.

Keywords

Education, Learning, LLMs, Reflective, Analytics Dashboard

ACM Reference Format:

Ka Hei Carrie Lau, Nađa Terzimehić, and Enkelejda Kasneci. 2026. PromptMirror: Visualizing LLM Use to Support STEM Student Reflection. In *Designing Interactive Systems Conference (DIS '26), June 13–17, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3800645.3812863

1 Introduction

Large language models (LLMs) have become increasingly embedded in university students' workflows, reshaping academic tasks such as brainstorming [43], summarizing literature [3], and drafting essays [3, 69]. Surveys and systematic reviews confirm widespread adoption and largely positive perceptions of tools such as ChatGPT, especially for productivity and writing support [4, 16, 67]. Yet students often misjudge their reliance, particularly under heavy workload and time pressure [1]. This misjudgement, and thus unreflective reliance on LLMs, can foster reduced academic performance [1, 33], superficial learning [16], diminished independent analytical skills [1], or integrity concerns such as plagiarism [22]. These risks highlight the need for a tool that helps students gauge and critically reflect on their LLM use [78].

Reflection is a well-established metacognitive practice that supports deeper learning, self-regulation, and critical thinking [12]. In Human-Computer Interaction (HCI), it is understood as critically assessing past, present, or anticipated experiences in order to reconsider and adapt behaviors, beliefs, and assumptions [10, 66, 68]. Researchers often operationalize reflection by confronting users with data or experiences that challenge existing practices and stimulate meaningful behavioral or cognitive change [10, 68]. Reflection has been extensively studied in personal informatics [19], digital wellbeing [72], and educational practice [35], yet it remains largely unexplored in students' everyday interactions with LLMs. Addressing this gap is critical, since habitual reliance on LLMs without reflective awareness risks undermining both learning depth and academic integrity [1, 22].

Current analytic dashboards for assessing LLM use remain mostly instructor-facing. For instance, Kim et al. [39] introduced PAD and Chen et al. [18] proposed StuGPTViz, both of which provide insights into students' AI interactions but primarily target educators. In contrast, several student-facing systems have explored supporting reflection and literacy in more specific contexts. Neshaei et al. [54] developed MindMate, a conversational agent to support reflective writing, Jin et al. [36] investigated utilizing chatbots embedded in learning analytics dashboards, and Dennison et al. [26] designed Prompty, a classroom tool that teaches prompt construction and critical evaluation. However, these are tied to specialized environments or emphasize skill training rather than fostering ongoing LLM use awareness. There remains a gap for a student-centric, reflective solution that supports awareness of everyday LLM interactions.

To address this gap, we draw from the reflection-by-design approach [10], and introduce and evaluate *PromptMirror*, an analytical dashboard that, to the best of our knowledge, is the first to allow students to upload their organic LLM conversation logs and visualize personalized analytics through four reflection-oriented components. These comprise usage statistics that surface temporal and frequency patterns of LLM interaction activity, sentiment analysis of both prompts and responses, intent analysis of prompt purposes, and word clouds that highlight thematic content. Each component was informed by two focus groups, one with experts and one with students, aligning with established reflective resources of temporal perspective and discovery [12]. To investigate the potential of PromptMirror to foster reflection, we pose the following research questions:

RQ1 Which metrics and design elements might support students' self-reflection on LLM usage within a personal analytics dashboard?

RQ2 How does PromptMirror influence students' awareness of their LLM usage and the depth of their reflection?

We first organized an expert focus group to identify metrics of visualizing LLM use that could foster reflection. Subsequently, we ran a student-facing focus group to establish common ground between expert and learner needs. Together, these sessions informed our answer to **RQ1** and provided the design rationale and design space that guided the development of our prototype, *PromptMirror*. To address **RQ2**, we conducted an online think-aloud study with $n = 20$ university students in two batches, who uploaded their organic, historic LLM conversation data and interacted with *PromptMirror*. Our findings provide preliminary evidence that PromptMirror may help students in recognizing their estimation gap on LLMs and cultivating greater self-awareness of LLM reliance. Moreover, qualitative analysis suggests that the dashboard fosters deeper and more reflective observations on their LLM use. Our contribution, as outlined by Wobbrock and Kientz [80], is twofold:

System We designed and implemented *PromptMirror*, a student-centric reflection tool that enables learners to upload their LLM interaction logs and reflect on their usage patterns through temporal, sentiment, intent, and thematic analyses.

Empirical We provide initial evidence that engaging with PromptMirror may raise students' awareness of their LLM use, resulting in an estimation gap, and foster more meaningful, reflective observations on their LLM usage.

2 Related Work

2.1 Perceptions of LLM Use in Education

The integration of LLMs in education has sparked widespread discussion in HCI and beyond. A recent meta-review [58] found that 15% of LLM-related publications at ACM's Conference on Human

Factors in Computing Systems (CHI) evolve around education. Together with a recent survey on the employment of LLMs for learning [74], which suggests that 86% of respondents have used LLMs for learning at some point, these findings highlight the growing relevance of addressing LLMs' role in educational and learning practices. Early commentaries emphasized both the benefits and risks of LLMs for and in learning. For example, Kasneci et al. [37] identified an opportune role of LLMs in writing, comprehension assistance, or critical thinking, while also warning of challenges such as overreliance, bias, misinformation, and data privacy. Lin [48] raises similar concerns for adult learners in self-directed learning, pointing to ambiguity in AI use policies and risks of cognitive dependency.

Empirical studies provide a more granular picture of how learners, particularly students, engage with LLMs. Sublime and Renna [70] surveyed 395 students aged 13 and above, finding that over 70% of their respondents adopted LLMs, with usage increasing by education level. However, only 20% to 50% actually revise the AI's output; this ability is more common among older students. This finding underscores the necessity of teaching AI literacy and reflective practices. Additional studies report growing dependency: Kim et al. [38] show that heavy users often delegate cognitive processes, such as reasoning and decision-making, to AI systems. In educational contexts, Abbas et al. [1] finds that greater workload and time pressure predict higher ChatGPT use, which in turn correlates with procrastination and lower academic performance. Kosmyna et al. [41] measure brain activity while using LLMs in essay writing and highlight *"potential cognitive costs"* such as diminishing students' intellectual engagement when compared to search engines and no technology use at all. Whereas Golding et al. [33] report that many students remain uncertain about using GenAI, influenced by perceptions of cheating in their decision to adopt such tools, another study by Combrinck and Loubser [22] demonstrates that students with high AI-detected scores often provide inaccurate or superficial information about their AI use. This reveals discrepancies between their declared and actual practices.

In response to this gap, our work explores how a visualization dashboard can support students in reflecting on their own LLM interaction patterns. By enabling students to see and reflect on their own LLM use, our aim is to initiate a discourse that will lead to a more mindful and effective integration of LLMs into learning practices.

2.2 Reflection in HCI

Reflection, broadly understood as the process of critical thinking about one's experiences to generate new understanding and inform future action, is foundational in both education and HCI research. Although there remains considerable ambiguity around the definition of reflection within HCI [10, 12], Donald Schön's notions of *reflection-in-action* and *reflection-on-action* [66] have been widely adopted in HCI [10] to conceptualize how users think *during* and *after* interaction with technology in the context of HCI [59]. Our developed system, PromptMirror, aggregates and visualizes historical LLM usage data, that is, after the LLM has been used, therefore aiming to encourage *reflection-on-action* taken.

Reflection's significance is particularly evident in the context of personal informatics, i.e., (self-)tracking of personal data, where reflection is considered a crucial stage in understanding and managing personal data [30, 46] towards raising awareness of one's own certain behaviors [12], resulting in new personal insights [19] and actionable steps [46]. To further break down the qualities of self-reflection, Fleck and Fitzpatrick [32]'s *"levels of reflection"* framework offers a useful lens. This framework proposes five levels of reflection: 1) R0 Description: Revisiting, 2) R1 Reflective Description: Revisiting with Explanation, 3) R2 Dialogic Reflection: Exploring Relationships, 4) R3 Transformative Reflection: Fundamental Change and 5) R4 Critical Reflection: Wider Implications. Choe et al. [19] use this framework to understand how people reflect on their past behaviors when they explore aggregated data with visualizations within their system Visualized Self. Their findings suggest that visually aggregated data can facilitate user insights reaching up to R2-level reflection, characterized by the exploration of relationships between data points and personal experiences.

Initially presumed as an automatic step if personal data has been simply made available [46], reflection has more recently been recognized as an essential process that needs deliberate design support and encouragement [9, 10, 12, 68]. In a recent literature and market app review, Bentvelzen et al. [12] constructed four design resources for interactive artifacts to foster reflection: temporal perspective, conversation, comparison and discovery. In other words, if an interactive artifact employs one or more of these design resources, they are about to spark reflection [12].

The HCI field has developed numerous systems in the past, reported to effectively facilitate reflection [12]. These systems address various application areas and employ diverse interaction techniques [12], one of which is fostering reflection through visual representation and exploration of data [15, 19, 75].

Visualized Self [19] is a web-based application designed to aggregate and visualize personal health data from multiple sources, independent of any specific application domain. It includes features such as multi-source data import, high-level data summaries, a trend exploration page, contextual data management tools, and a comparison view for deeper analysis. The study examines both the levels of reflection users engage in (as described by Fleck and Fitzpatrick [32]) and the types of insights users generate during visual exploration. Building on their prior work [20], the authors develop a taxonomy of personal insights, identifying frequent themes such as recalling external contexts, recognizing trends, and comparing data over time. Notably, the study introduces a new category of insight: value judgment. In this category, participants made emotional assessments of their data, which sometimes led to intentions to change their behavior.

We build on the methodology of Choe et al. [19], applying and extending their insight taxonomy to (1) analyze the types of personal insights elicited by our dashboard and (2) investigate how these insights relate to students' self-perceptions of LLM use and their subsequent behavioral intentions.

Other HCI systems are more specific for one application area. For example, *Reveal-it!* [76] prompts reflection on personal energy consumption levels in a public display setting. Users could upload their energy use data, which was subsequently visualized on the public display to enable social comparison between individual consumption patterns and neighborhood statistics – fostering awareness and public discourse through situated reflection.

Closer to our domain is the use of reflective technologies for understanding and regulating smartphone use, particularly screen time, through self-monitoring and visualization techniques [49, 62, 73]. For example, *meTime* is a lightweight widget that helps users reflect on their technology usage patterns. The system supported users in identifying distractions and improving their focus over time [79]. Another system, *ScreenLife*, tracks technology use – aka screen time – among multiple digital devices, showing *"the user when their devices were on and in active use"* [63].

The listed tools act as mirrors, showing individuals their data and fostering insights through visual representation in various areas of application, such as fitness, health, energy consumption or smartphone usage. However, despite the seemingly detrimental consequences for learning, similar to the effect of extensive technology use on digital wellbeing, the application of reflecting on LLM use has remained an under-explored area.

We first address this gap by collecting potentially useful LLM tracking metrics for reflection on LLM use. This is followed by the development and evaluation of PromptMirror, a visual dashboard that aggregates historic LLM usage data to raise awareness and encourage students to reflect on their LLM usage behavior.

2.3 Analytics and Visual Dashboards for Learning

Learning Analytics Dashboards (LADs) have long been designed to make invisible learning processes visible, thereby fostering students' reflection, self-awareness, and self-regulation. Early systems such as the Student Activity Meter (SAM) visualized time and activity traces to support awareness of effort and engagement [28, 34]. Subsequent reviews have mapped how LADs aggregate multiple indicators—such as grades, participation, time spent, and progression—into unified visualizations, typically derived from structured datasets such as *Learning Management System (LMS)* logs or *Massive Open Online Course (MOOC)* traces [65, 71].

Verbert et al. [77] trace the evolution of LADs from early awareness tools to multimodal and participatory approaches, and propose a research agenda that moves beyond usability to explicitly support reflection, behavior change, and learning impact. Building on this trajectory, later work has emphasized dashboards as tools for *reflection and sensemaking* rather than prescriptive feedback. For example, Ahn et al. [2] introduced journaling features to embed dashboards into teachers' reflective routines and cautioned against "signals of correctness," which risk undermining user sensemaking and turning dashboards into judgment tools.

Because LADs are rooted in institutional contexts and structured indicators (e.g., logins, submissions, grades), they are limited in capturing students' use of LLMs. Such use is *fluid* because it spans diverse goals and contexts, *exploratory* as students iteratively probe and reformulate prompts, and *affective* since it is entangled with emotions such as anxiety, guilt, or confidence, as shown in studies of the "secret use" of LLMs [82].

While recent systems have begun to operationalize LLM traces, they focus mainly on instructor- or expert-facing goals. For example, PAD analyzes prompts to inform teachers' understanding of writing processes [39], StuGPTViz visualizes student–ChatGPT conversation patterns for assessment [18], and PromptHive provides a platform for subject-matter experts to collaboratively refine prompts for educational content [13]. These works demonstrate the feasibility of analyzing and visualizing LLM interactions, but overlook students' own reflective needs.

Our work builds on the LAD tradition of supporting reflection but contributes a student-centered dashboard design space, tailored to LLM use. PromptMirror shifts the focus from institutional progress metrics toward *personal usage patterns, sentiment, and meta-reflection*, extending prior analytics approaches into a novel and increasingly critical domain of student–LLM interaction.

3 Metrics for LLM Use

To investigate which metrics bear the potential to support reflection on LLM usage, we conducted two focus groups: one Expert and one Student Focus Group. We ran the two focus groups as parallel efforts (although they took place at different times) to avoid findings from one group influencing the other and to prevent any biases from being introduced. The two focus groups should therefore be seen as complementary, rather than one iteratively informing the other. For both focus groups, we report the procedure, analysis and resulting metrics, next to design elements proposed within the Expert Focus Group. To address RQ1, we synthesize findings from both focus groups in a resulting design space for LLM reflection, as in Section 3.3.

3.1 Expert Focus Group

3.1.1 Method.

Procedure. We conducted a 70-minute expert focus group with four researchers. After a brief introduction and obtaining consent, we introduced the focus group's aim on exploring the role of self-reflection in the use of LLMs such as ChatGPT within educational contexts. We clarified our intention to identify relevant metrics and co-design elements to support student reflection on LLM use.

The focus group was structured into three discussion rounds. First, we asked participants to reflect individually and then collectively discuss the overall benefits and challenges associated with student self-reflection on LLM usage, focusing particularly on whether LLM over-reliance exists in students' usage, barriers to accurate self-awareness, and opportunities for enhancing reflective practices.

In the second round, we introduced theoretical frameworks from Baumer et al. [10], the Dimensions of Reflection Coding Scheme, and Bentvelzen et al. [12], the Four Design Resources for Technologies that Support Reflection, to guide the discussions. Participants individually brainstormed and then collaboratively identified key metrics, reflection indicators, and data-collection methods using a digital whiteboard (Miro) [52].

The third round involved a live demonstration of the very early PromptMirror prototype. Participants provided critical feedback, specifically discussing usability, clarity of analytics, potential cognitive overload, privacy considerations, and the prototype's overall effectiveness in facilitating student reflection. The entire session was recorded for transcription and subsequent qualitative analysis.

Participants. We recruited four experts (2M, 2F, age $M = 31.5$ years) with backgrounds in education, cognitive science, HCI, and privacy research to ensure a diverse and comprehensive exploration of metrics for reflection and LLM use. Table 1 provides a demographic overview of the participating experts.

3.1.2 Results. Overall, experts agreed on the usefulness of a tool that visualized LLM use metrics: *"Having insight into your data always makes room for reflection."* (E4) We further identified and categorized the expert suggestions into three main clusters of **metrics**: (1) usage and agency, (2) motivation and satisfaction, and (3) prompt evolution and emotional dimensions. Across these clusters, experts emphasized that reflection should move beyond simple usage counts toward indicators that capture deeper engagement, critical awareness, and affective experience.

Usage and Agency Experts emphasized that reflection requires distinguishing surface-level use from deeper learning activities. *"Agency is super important,"* proposing metrics that separate "low-level" tasks (e.g., extracting information) from "high-level" learning goals. (E3) They also stressed the need to reflect on potential biases in LLM outputs and their influence on attitudes. *"Younger people are often not aware of these biases and would need to be made aware before they could fully self-reflect."* (E3)

Motivation and Satisfaction Experts suggested metrics that capture why students engage with LLMs, how satisfied they feel, and how they compare to others. *"Why they chose to use an LLM in the first place, how satisfied they felt with its output, and whether its quality or usefulness encouraged them to adopt it for other tasks beyond the immediate assignment."* (E1)
At the same time, they questioned what such comparisons mean without a ground truth. *"Even if I see other people use it this much and I use it more or less, what does this information really tell me?"* (E2)
Others proposed extending reflection to broader domains such as environmental awareness. *"LLM use constantly uses up resources . . . people are not very aware of it when they use it."* (E2)

Prompt Evolution and Emotional Dimensions Experts proposed metrics that track how prompts evolve to reveal learning processes. *"Differences in wording or added information"* across iterations show how students refine their thinking. (E2, E4)
"Students may form an emotional attachment with ChatGPT that influences how they reflect on its use." (E4)

In addition, experts envisioned **design elements** that could shape how reflection tools support students. Their suggestions varied in required effort, guidance, and interaction style.

Explicit User Feedback Experts recommended lightweight feedback mechanisms, such as a satisfaction checkbox with Likert-scale answers. An optional open-ended follow-up like *"Why were you satisfied?"* could deepen reflection. (E1)

Personalization and Actionable Suggestions Experts stressed the importance of turning data into action. *"People are not satisfied if you only present them data and leave it to them to make sense out of it."* (E2) For example, frequent translation prompts could trigger a recommendation: *"If you're doing a lot of English–French translations, would you like to look into this free course?"* (E2)

Prompt Evolution and Reflection Experts highlighted the value of visualizing prompt iteration. *"How often [one] refined a prompt until [...] the final result"* can help students review their own progression. (E2)

Guidance versus Neutrality Experts debated whether dashboards should remain neutral or guide reflection more directly. *"What if it's just like a tool to get the shocking moment of, 'oh, I'm really spending much more time with it'?"* (E3) *"Guide them towards reflection in the least amount of guiding that you can, but still some kind of guiding."* (E1)

Based on the expert focus group, we developed four wireframe prototypes (Figure 2). These visualizations illustrate motivation tracking, satisfaction trends, and micro-reflection prompts, and represent directions for extending PromptMirror. Future work will implement and evaluate these features at scale.

3.2 Student Focus Group

3.2.1 Method.

Procedure. We ran the focus group online via Zoom for around 90 minutes. After a brief introduction to the project goals (*"talk about a dashboard, a visualization, to reflect on your large language model use, in particular for learning"*), we asked participants to brainstorm potentially useful metrics for tracking their own LLM use. Participants generated individual sticky notes for each metric in a shared Miro board, which they subsequently presented in a group setting to prompt discussion and clarification. Participants then prioritized the generated metrics using the MoSCoW method, a prioritization technique that categorizes items into Must have, Should have, Could have, and Won't have requirements [21].

We transcribed the focus group for qualitative analysis, with one author conducting a content analysis. The analyzing author then extrapolated the resulting metrics with the metrics' clusters that emerged within the Miro board and discussed their findings with a second author upon agreement.

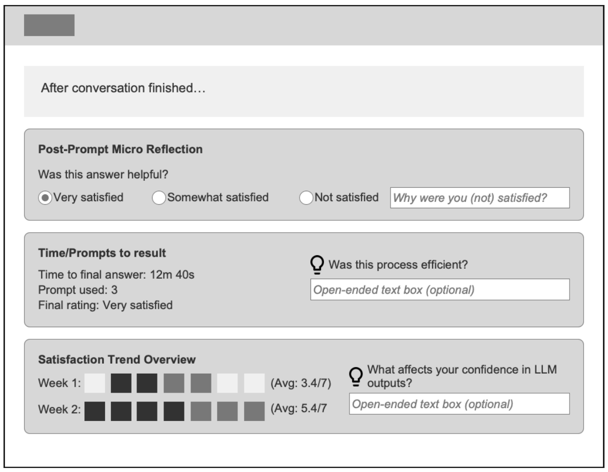
Participants. We recruited four students, aged 23-25 (2M, 2F), from a technical university setting (see Table 1). All participants were frequent LLM users, employing tools such as ChatGPT, Copilot or integrated search assistants (e.g., Gemini in Google search).

3.2.2 Resulting Metrics. We identified 14 distinct LLM use metrics categorized into three main clusters: 1) usage patterns and time allocation, 2) topics and tasks, and 3) quality and effectiveness of LLM interactions. The identified metrics are not always exclusive to their respective categories. For instance, participants wanted to see how much time they spent on a single LLM in total, as well as how much time they spent on each task or topic (intertwining the categories *time usage patterns* and *topics and tasks*).

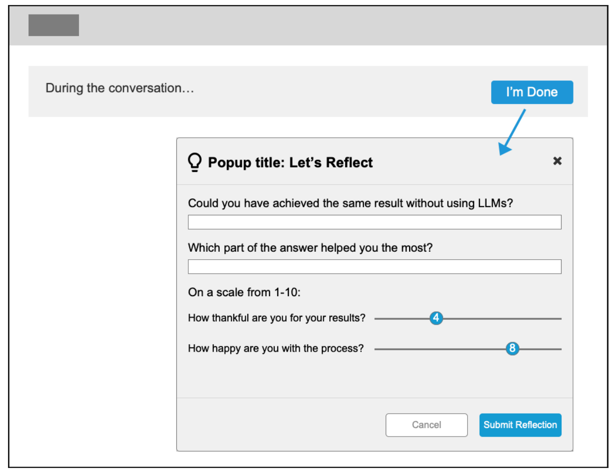
Beyond the three primary clusters, one participant raised the issue of environmental impact, pointing out that LLM usage may carry greater energy costs than alternative tools – thus wishing for a visual presentation of the fact in order to raise awareness. Per scope of this paper that aims to reflect individual LLM use patterns,

Table 1: Overview of Background of Focus Group Participants

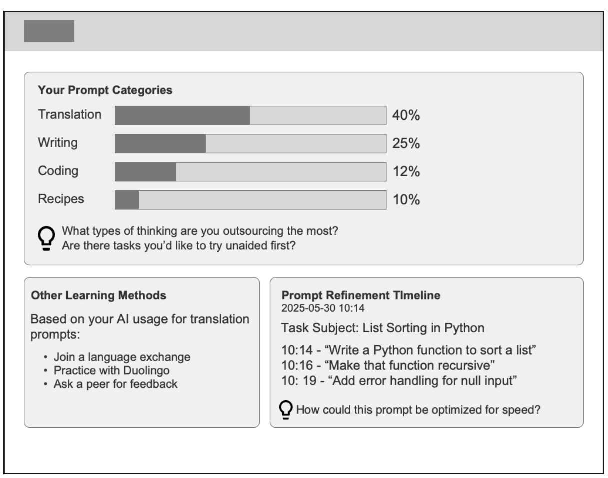
Expert Focus Group				Student Focus Group			
ID	Age	Gender	Expertise and Research Focus	ID	Age	Gender	Study Background
E1	34	M	Linguistics, Education researcher	S1	23	M	Computer Science
E2	26	W	Cognitive Science, Teaching	S2	25	W	Medical-Engineering
E3	35	W	HCI Researcher	S3	24	W	Human Factors Engineering
E4	31	M	Privacy, Smartphone Sensing	S4	23	M	Computer Science



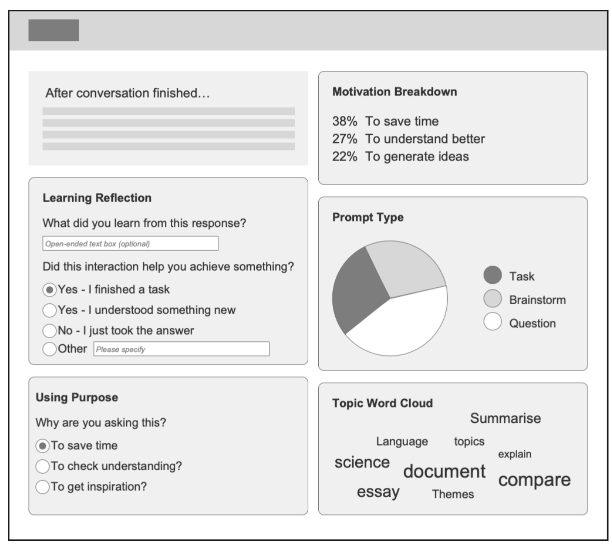
(a) Post-prompt micro-reflection (E1)



(b) Reflection-in-action (E2)



(c) Usage overview (E3)



(d) End-of-session dashboard (E4)

Figure 2: Four wireframe prototypes extending PromptMirror with reflection features, based on expert suggestions: (E1) post-prompt micro-reflection with satisfaction rating, efficiency tracking, and trends; (E2) reflection-in-action triggered by meta-questions on process and LLM reliance; (E3) usage overview with task categories, prompt refinement timeline, and alternative learning methods; (E4) end-of-session dashboard with motivation breakdown and prompt type distribution.

we illustrate the three main clusters and their specific metrics with corresponding quotes below. However, we acknowledge the importance of a wider societal impact of LLMs and the need to include societal metrics in subsequent versions of the visual dashboard.

Usage Patterns and Time Allocation All participants emphasized the importance of tracking either the individual and combined time spent interacting with LLMs (S1, S3 & S4), the frequency of opening/using LLMs (S1, S2), or the message exchange rate (S3). Participants subsequently wanted total usage time and session statistics visualized.
"What am I spending my time on, how much time I'm studying, how much time I am just asking random questions, kind of like screen time also for tracking. How well I'm spending my time basically." (S4)
For S2, this metric presents a proxy for their dependency over LLMs, as stated:
"Frequency of opening the [LLM] app [...] kind of reflects [...] my dependency on [the LLM] tools, I guess." (S2)
S3 also briefly discussed latent, unintentional AI use and the importance of tracking exposure time to those: *"Also, for example, for Google search, every time I Google something, I also get this LLM answer instead of clicking on different websites"*.

Topics and Tasks The second cluster of metrics evolved around categorizing the content and purpose of the interactions with LLMs. Students want to know about the topics they are asking LLMs about (crossed with time spent on each topic, as per S4) and the tasks they use LLMs for, depending on the different contexts – *"do I use it more in a professional way for university or my work or private?"* (S3)

Quality and Effectiveness of LLM Interactions Finally, all participants wanted some sort of metric to show whether they are satisfied with the LLM interaction or whether the LLM had helped them solve a problem. S4 acknowledged the difficulty of calculating such a metric:
"The percentage of helpful responses. So, how often do I have to correct ChatGPT? I imagine this is probably not super easy to measure because I also, when I get a wrong answer, I just stop asking it."
S1 suggested using the sentiment of chats, i.e., the amount of positive or negative messages, to quantify how successful the responses were or even adding a post-interaction pop-up for user feedback stating *"Was this chat actually helpful to you?"* (S1)

3.3 Resulting Design Space for LLM Reflection (RQ1)

To answer **RQ1**, we combined insights from the Expert Focus Group and Student Focus Group into a design space of reflection metrics and features (Figure 3). This space organizes participant suggestions into five categories for operationalizing reflection on LLM use.

Core Metrics of Usage Participants stressed the importance of tracking *time spent with LLMs*, the *type of task* (e.g., writing, coding, brainstorming), and the *frequency of prompts and openings*. They considered these "must have" baselines that help students identify intensive-use periods and recognize patterns of dependence.

Dependency Indicators Participants proposed metrics that reveal how much students rely on LLMs. Students suggested tracking how often they *correct or abandon LLM outputs*, treating this as a proxy for over-reliance when they accept results uncritically. Experts recommended comparing LLM activity with other forms of work (e.g., manual writing or other learning platforms) and highlighting shifts in reliance over time.

Quality and Affective Dimensions Experts pointed to the role of *emotional satisfaction* (e.g., frustration, confidence, relief), while students emphasized perceived *usefulness and productivity*. Sentiment-based measures can trigger meta-reflection by showing when and why LLMs feel supportive or overwhelming.

Thematic and Contextual Perspectives Participants valued organizing LLM use by *themes and context* rather than only by time. Students proposed clustering prompts into topics or visual word clouds, and distinguishing between professional and personal content, to reveal where they rely most on LLMs. They also wanted comparisons across peers or courses to identify whether heavy use related to specific assignments or practices. These thematic and contextual views situate reflection in broader educational and social settings, highlighting not just how much students use LLMs, but also *why and where* they matter most.

Reflection Prompts Experts, in particular, envisioned features that move beyond metrics to foster reflection. They suggested disclosing which AI models students used, showing examples of effective prompting, and asking reflective questions (e.g., "Could I have completed this task without the LLM?"). Such support aims to foster intentional engagement and guide students toward metacognitive reflection, rather than leaving them with raw usage data.

4 *PromptMirror*

4.1 Design Rationale

Guided by the design space, PromptMirror's implementation integrates four metrics that span usage, affective, task-level, and contextual perspectives (Figure 4). For the purposes of this study, we did not include all elements of the design space, particularly some stemming from the Expert Focus Group, as these require more effort to reflect on. These will be kept for future work.

We developed PromptMirror in two iterations (named *v1* and *v2* in the following). The metrics from Figure 4 were persistent in both versions, however we improved the classifier for the sentiment and intent analysis in a second iteration and resolved performance issues so participants could upload all their conversation data from up to three sources instead of ChatGPT only. Furthermore, in a second iteration, we added a *Filtering* option.

4.2 System Architecture

4.2.1 Implementation Details. PromptMirror is a web dashboard that helps students reflect on their LLM usage by analyzing their conversation histories. The system consists of two main parts. The

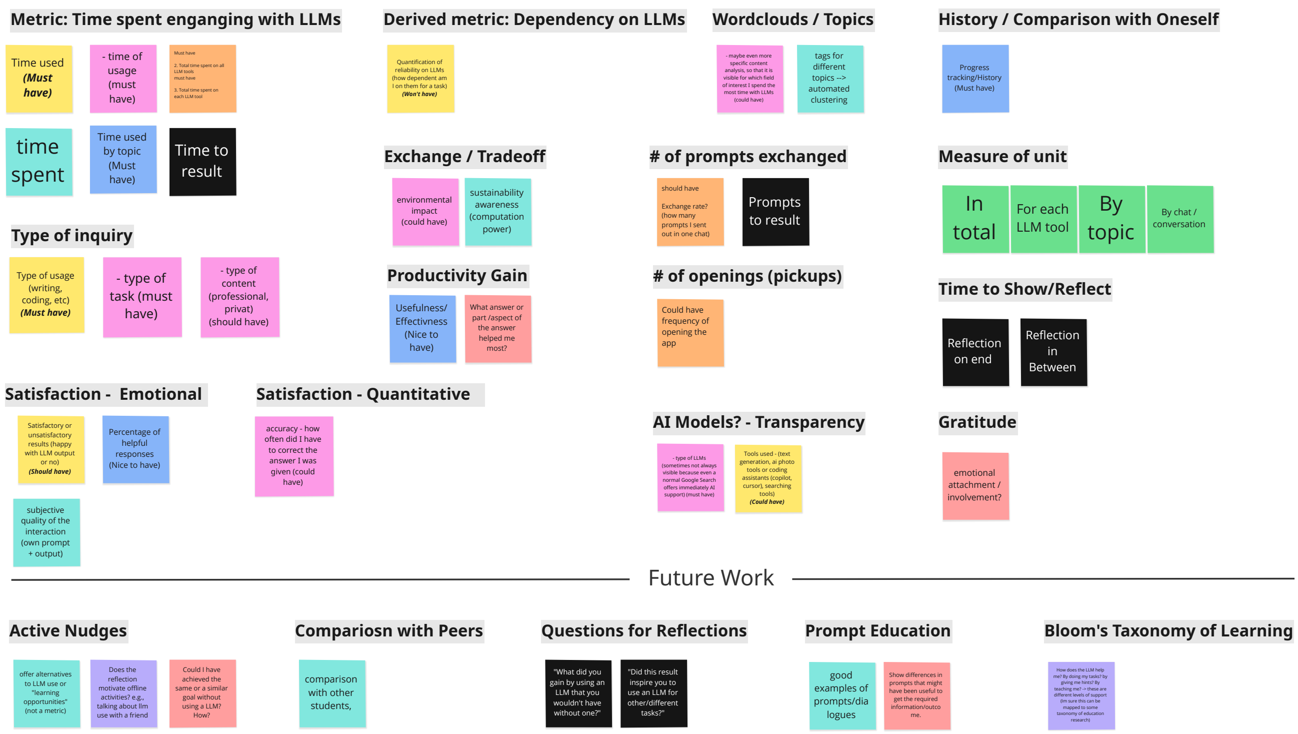


Figure 3: Consolidated design space of student-centric reflection metrics and features for PromptMirror, derived from the Expert and Student Focus Groups. Sticky notes represent proposed measures and ideas contributed by participants, clustered into different categories.

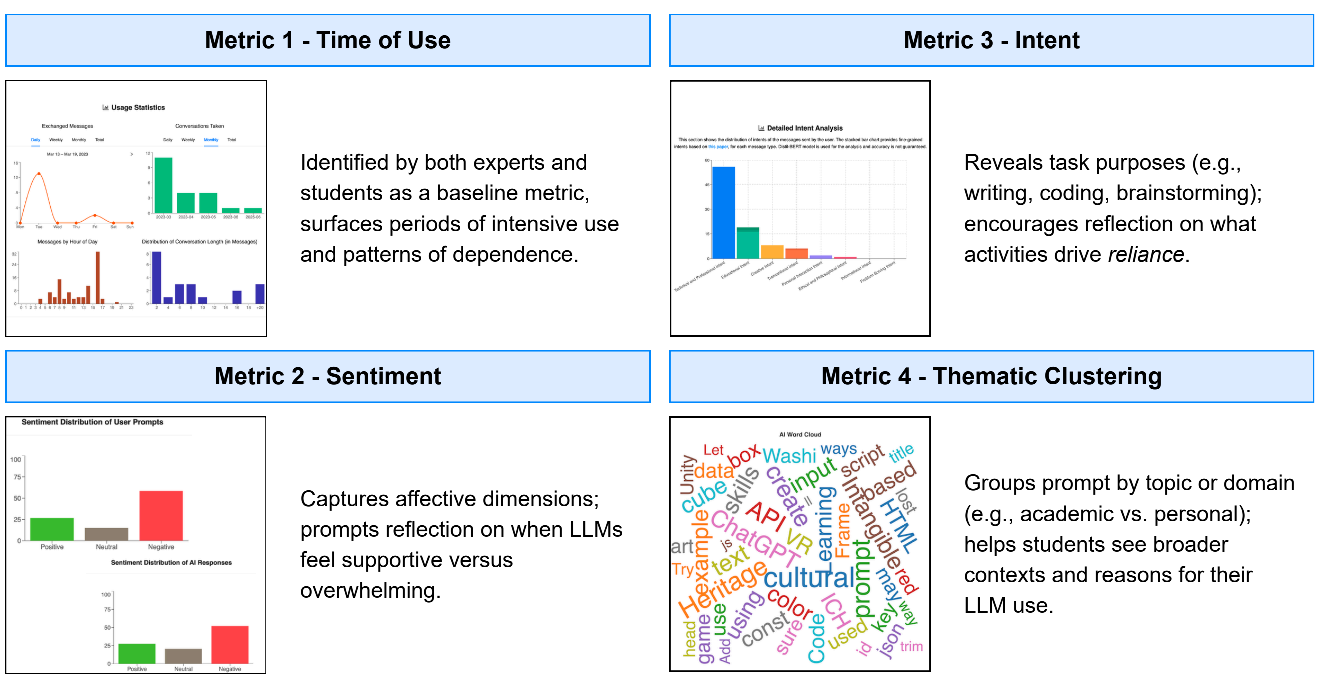


Figure 4: Design rationale for the four reflection metrics in PromptMirror, with illustrative screenshots.

backend, implemented in the .NET framework [51], provides web services that locally run a lightweight natural language processing model, DistilBERT [64], for sentiment and intent analysis. These models are converted into the Open Neural Network Exchange (ONNX) format [55] and executed with ONNX Runtime.

The frontend is built with the ReactJS library [50] and uses Ant Design [6] components for the UI. We used data visualization libraries such as Recharts [61] and D3 [24] for charts, and the react-d3-cloud library [81] to generate word clouds.

4.3 Features

Data Upload. Students can export their organic, historic conversations from ChatGPT (in v1) or up to three LLM services in v2 – ChatGPT, Gemini and Claude – as JSON and upload them to PromptMirror for a single analysis view. A date-range selector filters which conversations are analyzed before upload.

Usage Patterns. Messages are grouped by timestamp to compute daily/weekly/monthly counts, hour-of-day distributions, and conversation-length distributions (messages per conversation). Results render as line and bar charts.

Sentiment Analysis. The sentiment analysis feature has undergone two iterations. In both, each message is classified as positive, neutral, or negative using a fine-tuned DistilBERT. Text is normalized (lowercasing, punctuation removal) and tokenized to align with what the model was trained on. The model outputs two scores; we report the higher-confidence label and its confidence. However, in the first iteration that was presented to the first eight participants, the sentiment model tended to over-predict negative labels, especially in problem-solving exchanges where technical vocabulary may be misinterpreted as negative tone. This bias could have distorted the aggregate sentiment metrics, which is why we enhanced the DistilBERT fine-tuning process by providing classification examples. This produced more neutral message categorizations.

Intent Detection. We use a separate DistilBERT model again, and assign each message to one of several pedagogical intent categories, based on the taxonomy from previous related work Bodonhelyi et al. [14]. The model returns a single top class per message. Distributions are summarized as stacked bar charts.

Topic Extraction. Word clouds visualize frequent terms after normalization and stop-word removal. Term sizes use logarithmic scaling to reduce domination by very common words. Separate word clouds are shown for users and AI messages, respectively.

Visualization Toggles. Each visualization can be included or excluded — the system provides toggles that allow users to customize the visualization features according to their preferences. By default, all toggles are set to show visualizations.

Filtering (v2 only). Users can filter all metrics by LLM tool, model, date range, and intent, as shown in Figure 5.

Privacy-Preserving Design. The tool processes conversation text on demand, not persisting it server-side. Basic aggregations were computed client-side; the server produced sentiment and intent labels in memory and returned only derived results. The client discards uploaded data on reload or session end. While user/project metadata support access control, raw conversation content itself was not stored.

5 User Study: Think-Aloud Evaluation

To investigate whether *PromptMirror* can foster students' reflection on their use of LLMs, we derived **RQ2**: "How does PromptMirror influence students' awareness of their LLM usage and the depth of their reflection?" We addressed this question through an online think-aloud evaluation of the tool. We conducted the study in two rounds. In the first round, we recruited 8 participants and used an early version of the system that exhibited some issues. Based on participant feedback, we improved the system – most notably addressing performance, as well as sentiment and intent classification problems – and added a filtering option next to uploading LLM use data from multiple sources ChatGPT (v1 and v2), Gemini, Claude (v2 only). We then ran a second round with 12 participants using the revised system. In both rounds, the study procedure remained unchanged, and we report results jointly across all 20 participants.

5.1 Study Design & Procedure

We conducted all think-aloud sessions remotely via Zoom. The study procedure is presented in Figure 6.

After registering for the study, participants were sent a welcome email that asked to download their organic, historical LLM use data to reflect participants' actual, longitudinal use. The email contained instructions on how to export their LLM data from the respective LLM services' official website, using the data export function for ChatGPT [17] and Claude [7], as well as a tutorial for Gemini exports (see Appendix B). The exported files were in a JSON format, containing a tree-like structure of users' previous conversations and messages with the LLMs. Participants kept the JSON-files on their machines locally for the study and did not send us their raw data in any way. One to two days prior to their scheduled session, participants were sent their participant ID (PID) and a link to a demographic survey administered through Qualtrics [60].

At the beginning of each study session, we welcomed participants and obtained digitally signed and verbally confirmed informed consent for both video and screen recording. Before starting the think-aloud activity, we spent approximately two minutes onboarding participants. Following the practice from Ericsson and Simon [31], we explained what the think-aloud method entails, made clear that there were no right or wrong answers, and reminded them of the overall study goal. Once participants confirmed that they understood the task, we provided access to the *PromptMirror* dashboard and supported them in registering and uploading their LLM usage data. However, we did not introduce the *PromptMirror* tool beforehand nor explain its functionalities and features, in order to grasp participants' initial reactions to their LLM usage statistics and not to shape their PromptMirror usage patterns. Participants imported their previously exported LLM use data (only conversation files) from their local machine and were reassured that all data was processed on and by our university's servers (i.e., no API calls to external services) and deleted at the end of the session.

During the think-aloud session, participants explored their personal analytics (e.g., usage frequency, temporal patterns, spikes, and intent analysis) while verbalizing their thoughts. The facilitator

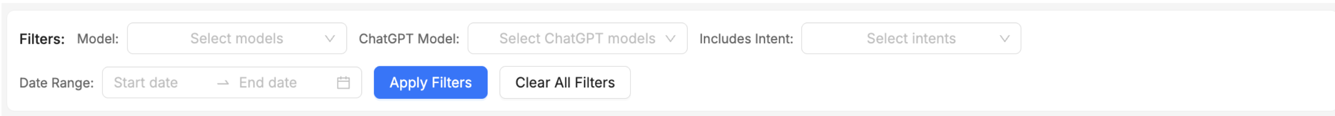


Figure 5: The Filtering feature at the top of the page.

occasionally asked clarifying questions or provided prompts to sustain reflection (e.g.,*"What does this trend or number tell you?"*, *"Are there patterns you notice here?"*).

Following the exploration, a short debriefing interview invited participants to summarize their impressions. The facilitator also asked questions about perceived insights and concerns (e.g.,*"What do you like most about the dashboard?"*, *"Do you have any privacy concerns when using such a dashboard?"*). After the debrief, the recording was stopped, and participants completed a technology-supported reflection inventory (TSRI) [11] via Qualtrics.

Each session lasted approximately 30–45 minutes, depending on the depth of discussion. All sessions were video-recorded for subsequent transcription and analysis.

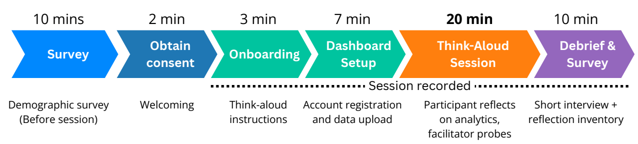


Figure 6: Session flow and timing of the think-aloud study. The demographic survey was conducted before the session, while the subsequent phases composed the 45-minute recorded session.

5.2 Participants

In the first study iteration, we recruited 10 students. One did not attend, and another withdrew due to concerns about sharing sensitive personal data. For the second iteration, we sought to recruit 12 additional participants. Of the 12 students who initially expressed interest, four dropped out after receiving instructions to download their LLM data: one cited privacy concerns, one incompatibility with our data collection process (e.g., using other LLM services such as Perplexity) and two students were a no-show. We were able to recruit an additional four participants, bringing the total pool from both study batches to $N = 20$.

Participants were required to be (i) at least 18 years old, (ii) currently enrolled in a university, and (iii) active users of ChatGPT, Gemini or Claude. Recruitment was conducted through convenience sampling. To address privacy concerns, we explicitly informed participants that no uploaded data would be stored. Each participant was compensated with a 15€(v1) or 20€(v2) gift-card.

The demographic survey collected both demographic information and participants' prior LLM use experience. Table 2 reports the key demographic and usage characteristics relevant for contextualizing our sample.

Participants' ages ranged from 21 to 38 years ($M = 25.8$, $SD = 3.85$). The sample included ten male and ten female participants, all of whom reported using LLMs for more than one year. Correspondingly, all participants used LLMs for a variety of tasks, with *writing* being not selected only in two participants.

Table 2: Participant demographics and LLM use characteristics ($N = 20$). All participants had more than one year of LLM experience; six reported paying for access (typically €10–50/month, one >€50).

Demographics			LLM Use Behavior		
ID	Age	Gender	Use Frequency	Paid Version	Regular LLMs
P1	29	M	Daily	Yes	ChatGPT
P2	21	F	Daily	Yes	ChatGPT, Copilot
P3	25	M	Daily	Yes	ChatGPT
P4	29	M	4–6/week	No	ChatGPT, Claude
P5	22	F	Daily	No	ChatGPT
P6	25	M	4–6/week	No	ChatGPT, Other
P7	30	M	Daily	No	ChatGPT, Grok
P8	26	F	4–6/week	No	ChatGPT, Claude
P9	30	F	Daily	Yes	ChatGPT, Gemini
P10	23	F	Daily	No	ChatGPT, Gemini
P11	22	M	Daily	Yes	ChatGPT, Gemini, Claude, Copilot, DeepSeek
P12	23	F	Daily	No	ChatGPT, Gemini
P13	26	F	Daily	No	ChatGPT, Claude
P14	25	M	Daily	Yes	ChatGPT, Gemini, Copilot
P15	24	M	2–3/week	No	Gemini
P16	24	F	4–6/week	No	ChatGPT, Gemini, Claude, Copilot
P17	23	F	4–6/week	No	ChatGPT, Gemini
P18	27	F	Daily	No	ChatGPT, Gemini
P19	23	M	Daily	No	ChatGPT, Gemini, Claude
P20	38	M	Daily	No	Gemini

5.3 Analysis

We analyzed questionnaire responses, audio recordings of think-aloud sessions, and screen recordings of participants' interactions. Audio recordings were transcribed using Condens [23], and questionnaire responses were exported and cleaned prior to analysis.

For the think-aloud sessions, we performed a content analysis as follows: two researchers first independently open-coded three think-aloud sessions using a mixed approach (i.e., both deductive and inductive coding). For deductive coding, we followed the codes developed by Choe et al. [19] on the types of insights that emerge with visual data exploration. Within this step, we discovered four new codes compared to the work of Choe et al. [19].

At the same time, we inductively coded participants' emerging LLM use patterns and opinions on (the use of) LLMs. After coding the three sessions, the two researchers compared their findings in a co-located session, discussing discrepant observations upon agreement resulting in a shared code structure. As per [8], we decided not to calculate inter-rater reliability because the units of observation in the think-aloud sessions were not comparable enough to apply standard measures. We then proceeded by splitting up the remaining sessions among the two researchers, who independently coded the sessions based on the developed code structure. We repeated the same procedure for the second study run, now coding two recordings separately and comparing our findings. This process yielded no new codes.

The final step included a joint refining of the codes into higher-level themes. These themes describe how participants engaged with *PromptMirror* and how they reflected on their LLM use, resulting in four overarching themes that we present in the following section. We present the results of the post-study TSRI scale (Figure 7) as complement to the qualitative findings.

6 Results

Our mixed coding approach first adapts Choe et al.'s insight taxonomy for data-driven reflection [19]. We used this taxonomy to inform the emerging reflection patterns. This approach resulted in the identification of four codes distinct to the taxonomy of Choe et al. [19], namely: *Holistic Judgment*, *Assumption*, *Consequence* and *Meta-Reflection* (see Table 3). Our inductive coding resulted in Table 4, presenting a taxonomy with LLM-specific reflection categories grounded in our focus groups and think-aloud evaluations.

In the following sections, we first present our findings on the reflection processes and the types of insights participants gained through visual exploration of their usage data, building on prior work from Choe et al. [19]. These reflections often led to increased self-awareness and prompted reconsideration of personal LLM-use habits. We then focus on findings specific to learning with LLMs, as detailed in Table 4, where the visual dashboard served participants as a useful aid to notice patterns in their LLM interactions. While not prompting immediate change, the visualization made abstract behaviors, such as frequency, timing, or reliance on LLMs, more concrete, thereby offering participants a starting point for reflecting on and potentially adapting their use over time, as the results from the TSRI suggest in Figure 7.

6.1 Reflection Patterns

Our participants employed similar reflection patterns to those found by Choe et al. [19]. To the best of our knowledge, PromptMirror is the first tool to visualize LLM use patterns in a visual dashboard. As such, our participants were for the first time confronted with their LLM use, which resulted in interesting observations.

After initial upload, all participants started by simply naming values and references in the upper part of PromptMirror, that presents the exchanged messages and started conversations (echoing the basic level of reflection, R0 [32]). However, as soon as they moved further down to the temporal distribution of messages and conversations, participants would identify peaks and start discussing whether these make sense and how they overlap with their own awareness, expressing either confirmation (*"Intent analysis, technical and professional intent. Yeah, that mostly checks out"* (P15)) or surprise about their LLM use (*"for some days I [...] haven't used [AI], which is quite a surprise to me. Maybe I'm on a different platform, on Claude or something."* (P19)). The majority of participants then strove to provide external context to their interaction patterns and produce correlations between the data, identify temporal trends or explain the causes for the data variability or contradictions opposed to their impression, as with P14:

"I feel like the more recent we get, the smaller the breaks are because I feel like I use ChatGPT daily." (P14)

Some participants felt invited to judge their values or overall data, both positively and negatively:

"Technical, professional intent. Yeah, that makes a lot of sense for me actually, because I use it daily for like coding and other questions in that sense. Also, it's good to see for me that I don't really share a lot of personal information." (P11)

These patterns correspond to what Fleck and Fitzpatrick [32] characterize as R1 and R2 – descriptive and dialogic reflection – in which individuals return to earlier experiences or data, using contextual cues to make sense of or rationalize particular findings and subsequently establish relationships between data or context.

Although our think-aloud protocol did not include questions about the benefits and challenges of LLMs for learning, many participants themselves, by providing context to the detailed intent analysis, initiated critical evaluations of the influence of LLMs on their learning processes, problem-solving abilities, and their overall approach to study and work tasks. Several participants expressed actionable change points (*"I might like reconsider whether I want to subscribe to [the LLM service] or not."*), demonstrating transformative intentions (i.e., R3 levels of reflection [32]).

Finally, in two instances, P5 and P9 referred to the environmental impact of excessive LLM use (*"waste a lot of resources [...], like energy [with] high amounts of queries"*) (P5), displaying a level of reflection equivalent to R4, i.e. *"critical reflection with wider implications"* [32].

6.2 (Un)Awareness of Interaction Patterns

While visually exploring their LLM use data, participants articulated their observations and realizations on *how* and *how much* they use ChatGPT. This led to insights about their reliance and the nature of their interactions with the LLM, next to the emerging desired behaviors around LLM interaction patterns.

Some participants recognized the sheer volume of messages or conversations exchanged with ChatGPT as surprising, with P5 stating *"what shocked me most was the number of queries I've used it for"* (P5) – the participant found having 140 conversations in three months.

Resembling the thought, P2 expressed *"'I didn't expect it to be so high, but it seems it actually seems true.'"* Interestingly, P2 at first seemed *"quite happy with how I currently use ChatGPT. I think it's helpful, but at the same time, we don't overuse it"*, however leaning towards a need for a behavior change towards the end of the think-aloud session: *"My observation is that maybe I should limit my usage a little when it comes to work is to, you know, not to follow into the, not to end up in the situation where ChatGPT does all the thinking for me and I just copy and correct its answers"*. Although the dashboard reportedly reflected their expectations (judged as being *"on the right way"*), P6 also reported a need to reduce their educational intent due to ChatGPT's seeming disability to help with more complex study topics in higher semesters. P4 highlighted the dashboard's usefulness to observe *"how dependent [they] are [...] on the large language models"* and is keen to see *"how much [they] have changed in [the last] 2 years"* regarding their usage patterns,

Table 3: New data-driven insight types (compared to Choe et al. [19]) used to analyze ChatGPT usage reflections.

Category	Description	Example Quote
Holistic judgment	Evaluation of overall data rather than single data points	*"My dashboard really represents, I'm using it for technical appliances and not like for random conversations."* (P3)
Consequence	(Prospective) action on the reflected data	*"Maybe I should delete [the conversations] after the study."* (P2)
Assumption	What would have happened, if the external context was different	*"So maybe if I use this thinking mode, maybe I use it more often also when I have more complex things and then these complex things, they have more emotions."* (P14)
Meta-Reflection	Reflect on how the tool itself is helping in reflection	*"So maybe for AI usage [the dashboard] is useful also because we are [using AI] more and more. So maybe it's going to addiction. So like we can control here. That's what I like."* (P18)

Table 4: Consolidated categories of student reflections on LLM use.

Category	Subcategory	Description	Example Quote
LLM Use Practices	Conversation management	Practices of organizing interactions, e.g., deleting, keeping, or splitting chats	*"I usually delete conversations with ChatGPT that are long or irrelevant, but keep ones I might return to."* (P2)
	Task-specific use	Using LLMs for academic, non-academic, or technical tasks	*"When writing job applications, I relied heavily on ChatGPT for text reformulation."* (P2)
	Role assignment	Naming or framing the model for specific functions	*"I named it "Mail Generator" so I could quickly draft professional emails."* (P6)
	User habits	Prompting styles or rituals	*"For some reason I always add "please" in my prompts."* (P3)
Opinions on AI	Trust and reliability	Perceptions of accuracy, hallucinations, and source fabrication	*"I do not usually trust ChatGPT to do my problem solving because it makes up sources."* (P2)
	Persona perceptions	Views on ChatGPT's tone or style	*"It is sometimes too friendly; I would prefer a more objective style."* (P2)
Coping Mechanisms	Prompt engineering	Learning how to prompt for optimal results	*"[I learned to] make my prompts more specific, almost perfecting them"* (P6)
	Alternative resources	Preference for traditional sources alongside ChatGPT	*"I can always go back to books for deeper understanding."* (P6)
	Privacy-preserving practices	Avoiding personal data disclosure	*"I never use real names or addresses; I replace them with placeholders."* (P6)
	Liability awareness	Keeping records for accountability or protection	*"I save prompts in case I need to show what was asked later."* (P5)

with P15 exercising a similar thought *"to see how one's use develops over time to see if you're becoming totally addicted or something."*

P11, with their 1572 conversations exchanged with ChatGPT, explained their surprise by making parallels to human conversations: *"Because that's more than 1/2 of probably most humans doing that entire time. [...] Also because you can't really enumerate how many conversations we've had with humans because there's no real counter to it. And also [...] it's not that big of a like a hassle to click new conversation rather than starting a conversation with humans."*

By contrast, P10 was surprised by the low number of messages and the short activity streak (i.e. the number of consecutive days of use): *"I'm actually surprised that it's just 14 days. [...] because I consider myself a regular user. [...] maybe weekends spoil the streak."*

The same goes for P18 and P13, who diversify their accounts or use LLMs when they are not logged in. This could explain why they have a lower number of interactions than they expected:

"I have my work which I use a lot because I also have premium in there and I have my personal one which is this one which I uploaded which I didn't use in a while." (P13)

Although P11 and P12 expressed surprise about the number of messages exchanged and conversations started, they were quite content with the *way* of their usage – mostly for technical and educational purposes and *not* for personal and emotional support. Other participants, such as Participants P1, P7, P8, P15 or P16, got their holistic use expectations confirmed – both for the usage numbers, as well as the use contexts of LLMs, which were for technical (i.e., coding) and educational intents mostly.

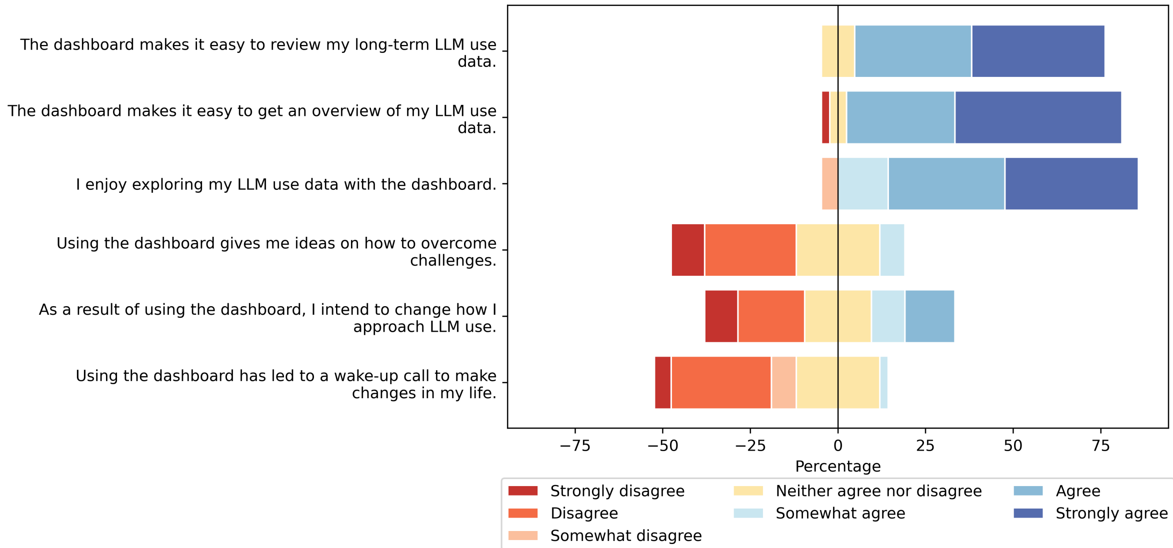


Figure 7: Results of the TSRI.

The listed participants centered their reflections on rather instrumental LLM use patterns. P1 focused on the cost-benefit analysis of using LLMs and the practical value of PromptMirror on making subscription decision: *"Is it really worth it that I subscribe to the premium version? Then I can take a look here and like decide, oh, like I'm using this a lot for technical and professional intent. Maybe, probably it's going to be like worth it to subscribe to the services"* (P1).

Participants additionally recognized the diverse range of topics for which they employ ChatGPT – from everyday questions like car functions (P2), medical reports, job application letters or interview preparations (e.g., P7, P9, P14) or grammar correction (e.g., P7, P8, P9) to more complex code logs debugging (e.g., P3, P11, P15, P20) or academic studies (all participants). The conversation topic often led to an additional discussion on the wide variability in conversation length. These range from *"really short conversations for quick, random questions"* (P5) to *"20 plus"* messages for complex tasks, such as debugging code or working on a thesis (P7). This pattern is also consistent for studying or programming, where participants expressed the need to stay on one topic for an extended period of time (P2, P8, P11).

Some participants additionally commented their recent increase of use, with a more skeptical approach at the beginning, but now embracing the technology, as with P15 and P17, respectively:

"[It] seems I'm using it more and more, at least so far." (P15)

"in the beginning I didn't really trust ChatGPT because I was, I was used to use Google. But in 2025, I used it a lot because it's like maybe it improved or started giving better information." (P17)

6.3 Reflection on Learning Strategies

Our general observation of all participants is that they use LLMs for learning and work purposes to a significant extent – as a highly adaptable tutor – reflected in the high number of educational intents presented within the dashboard.

Participants acknowledged both the benefits and the challenges of LLMs for learning, drawn from their own experiences. P3's statement echoes our observation: *"I'm using it almost exclusively for learning, analyzing [code] logs [...] it really helps because if [...] you have like a log that is really long, [...] it can easily spot the issues or whatever."* (P3) The majority of participants mentioned significant use during exam periods. Participants perceived LLMs beneficial for thesis writing and programming, concept explaining, summarization and data organization, brainstorming and code generation or language learning.

A prominent concern by P2 is that ChatGPT can lead to *"laziness"*, in particular for programming, admitting that they *"don't remember when the last time was that I actually wrote the code from scratch by myself"* instead using ChatGPT for blueprints. They highlighted this dependency by recalling a computer-based exam requiring code from scratch that became *"a little bit more difficult than in pre-ChatGPT times"*. At first, P5 viewed using ChatGPT as a helpful *"baseline"*, explaining that it *"points me in a direction, for example, by suggesting a library. From there, I can go to Google or the documentation to understand how the function works, building on what ChatGPT has provided."* By the end of the session, however, P5 expressed opposing views for their learning processes and drew parallels to wider societal implications: *"Maybe [...] I'm a little bit too dependent on ChatGPT sometimes to be like my first point of interaction towards solving a problem, which I don't know if it's good because I also know that ChatGPT queries like waste a lot of resources [... energy-wise]."*

Initially, P4 used LLMs daily, but their usage has since shifted specifically for programming-related tasks: by reverting to traditional methods of self-exploration like executing a Google search or reading through Stackoverflow. This shift stems from a concern that they *"didn't learn that much"* from problems solved by LLMs. Despite this, P4 finds LLMs particularly useful for learning languages (e.g., grammar corrections, translations or synonym understandings). Similarly, P18 drew parallels to pre-LLM times: *"So now for everything, I think they're going to do AI and asking everything even without thinking. So in my opinion, [...] AI is useful, but in somehow it's also reduced the ability of thinking and maybe some basic abilities. If we are just using AI, it's I mean like maybe I can not do any more."* P20 commented on their observation on how Google-ing has changed since the emergence of AI: *"when I really needed information, I go direct to Google that in any ways you are using Gemini somehow, right? Because when you read, I just read now the the that AI answer on the top. (P20)"*

In our first study run, the sentiment analysis tended to categorize messages as mostly negative, which however lead to two interesting observations: first, some participants nonetheless tried to find sense in the categorization by reflecting on their own behaviors and how these might have produced the negative sentiment; second, by sharing their past observation that ChatGPT is overly friendly (*"prioritizing user happiness over objective answers"*, P2), P2, P6 and P8 expressed their desire for a more objective, *"raw"* (P6) and critical feedback from ChatGPT: *"Normally they try to be neutral or positive like in this sense, which is sometimes not good because I would need a critical evaluation of some things, even though it would be negative. But a lot of times ChatGPT doesn't do it at all."* (P8) For P6, ChatGPT is an *"instrument, a working instrument [...] [like] a hammer or screwdriver"*.

There exists a strong awareness of LLMs' limitations and fallibility among our participants' pool, evident in Participants P12 and P15 were skeptical about the LLMs' hallucination property, being cautious not to use it for *"informational problems. They might hallucinate, they might have a lot of misleading information because we don't know where the sources are from. So I really don't use it for this. And when I have a something informational, I just go to Google directly and look for more reliable sources."* (P12) P15 explained how the insufficient amount of available training data for LLMs around certain topics has presented a true problem for them: *"I use Rust which is growing and quite popular but nowhere near as much data I guess, and also maybe more complex to write than something like Python. But yeah, I found it was just easier not to use it instead of having to go back and then recheck and rewrite."* (P15) P2's statement that *"ChatGPT is not a single source of truth for everything. It makes mistakes and it makes a lot of mistake."*, makes it all the more important to take a critical approach to the output of LLMs for learning purposes, as per P2.

P11 was much more optimistic, praising the capability of AI to teach: *"you learn a lot more through AI than you learn through other means, which I think is really interesting. And the responses are not shorter, but the problems are getting longer, meaning the answers must be more specified and the responses are more neutral as it just."* (P11) The same participant expressed how they use AI for all their learning, correlating their inactivity streak to the only times they have not been learning or working:

> *"My last longest inactivity streak went in July. That was actually the only time ever since that I've not had my laptop with me because I usually don't just chat with my phone."* (P11)

6.4 Student Adaptation and Coping Mechanisms

In a reflection sparked by PromptMirror, our participants delved into the deliberate strategies they use to manage their LLM interactions, optimize their learning performance, and mitigate any perceived negative impacts.

Interestingly, our first two participants P1 and P2, as well as P12, expressed following a pattern of curating/deleting *"most of [the conversations] once the question is completely answered"* if they *"won't meet this conversation in the future"* (P2) – in order to *"prevent [the LLM] from being biased toward a specific subject"* (P1) or out of *"privacy reasons"* (P12). With a similar goal of maintaining context and preventing the LLM to overfitting responses – i.e., to *"[think] much much much longer"* – P6 stated segmenting their conversations, treating each dialogue to practically fill out one task. In contrast, P11 never deletes anything but rather starts a new conversation for each task and copies prompts that worked beforehand.

Prompt engineering emerged as a key coping strategy – participants learn to *"make my prompts more specific, almost perfecting them"* (P5). P6 employed a specific system prompt to encourage a *"skeptical questioning approach"* to receive more objective answers.

P4, P7 and P8 described exhibiting a strategic approach to diversifying their LLM toolset for different tasks – Grok for philosophical questions, as per P7 and Claude for programming tasks as per P4 and P8. In our second study run, many participants used more than one LLM – as Gemini was not long before giving free yearly Pro subscriptions for students. P16 explains how they diversify their toolset as follows: *"Claude reasoning I really like. So I usually trust that more for like academic stuff than ChatGPT. I used to use it a lot until I discovered like Claude and Gemini, which I prefer."*

Privacy concerns lead users to anonymize personal data or to be very careful about not entering them at all. P8 consciously avoids *"giving the names, the real names of persons"* or *"real mail addresses"* and *"depersonalizing my messages"* due to past *"leaks of ChatGPT"*. On the other hand, P14 wondered how the LLM knew their name, which was prominent in the word cloud representation. P9 exercised no true concern, as eminent in their statement: *"but some of my friends are telling me that no, you shouldn't upload your own selfies. You don't know how it like what could have happened in the future, so they don't do it at all. But it's a personal choice, so I do it."* (P9)

Some participants actively monitor and correct LLM output when in doubt of LLM hallucinations. P7 recalled to *"see through the code, read through the code myself and point to the exact specific parts where it had to make corrections"* or if an answer *"is completely off, I would just do those same steps by myself"* (P5). P2 relies on their *"enough, enough knowledge to see that [the response is] OK."*

6.5 Meta-Reflection and Dashboard Improvements

Some of our participants exhibited self-reflection and reported on the dashboard's usefulness for this purpose, which we refer to as *meta-reflection.* Alongside meta-reflection, we present participants' suggestions for improving PromptMirror's features to enhance the dashboard's reflective capabilities.

Participants valued the dashboard as it provided *"useful information"* to help them *"make a judgment"* about the utility and dependency of their LLM use. While some find it merely *"interesting to see"* (P3) rather than directly useful for changing behavior, others appreciate that it *"confirms what I expected"* and shows them they're *"on the right way"* (P6). According to the TSRI results (see Figure 7), the majority of participants found exploring their LLM use with PromptMirror easy and enjoyable. Whereas participants see no value of PromptMirror (in its current form and metrics) in overcoming challenges or making changes in their life, some participants intend to change how they approach LLM use in the future: either by reducing LLM use, or by using the dashboard to optimize their LLM use, as per P9: *"Using AI better would definitely help me out and it is my main motivation."*

Regarding dashboard improvements, many participants expressed interest in peer comparison of usage, to *"upload your own [data] and someone else's and just have it directly give you the... comparison"* (P3). In fact, a third-person effect [25] is evident in some participants' reflections, where they observe or assume certain LLM use patterns more strongly in others than themselves. For example, P3 speculated on their partner's use, expecting *"a lot more conversations, but a lot, a lot smaller depth for each conversation"* compared to their own longer, more in-depth conversations related to deep work. P2 also displayed this effect by noting that *"many people, especially many young people, use ChatGPT Kind of as a psychologist"*, but explicitly stated, that they *"don't rely on ChatGPT for these purposes"*.

Participants wished for more detailed and comparable usage statistics. P1 suggests *"recap statistics"* similar to Grammarly, showing metrics such as *"how many documents you have... corrected... how many mistakes you faced, how many words you wrote"*. This would also help them assess *"how much I'm using them"* (P1) to justify, for example, subscription costs.

The sentiment analysis elicited surprise and contradiction. Some participants questioned its accuracy, noting that debugging questions might be classified as negative even when their overall interaction is neutral or positive (P5). P3, who *"always uses please in my prompts"*, was surprised when the sentiment analysis did not reflect positive interaction. Several participants mentioned wanting to see examples of classifications and snippets of their raw data showing how it has been classified, as P9 echoes:

> *"sentiment analysis I would like to have more details on that because it just because it confused me. [...] I feel like that should be a reason for me to use that dashboard, but if I cannot like get the reason why I'm seeing it or [...] if I don't know how to react to that analysis, I don't think it would be helpful. [...] again [if] I'm a negative prompt user and if I know the reason and if I think that oh maybe being negative results AI to response in this way, I might change my behavior."* (P9)

Privacy is a significant consideration for dashboard improvements. Participants appreciated that the dashboard *"doesn't [save] any personal data"* and has *"anonymization in place"* to prevent private conversations from being *"revealed or analyzed by third parties"*. P2 suggested to integrate the dashboard directly into the ChatGPT interface, similar to phone screen time statistics, making usage analysis easily accessible at its source.

Some participants expressed a wish to improve their LLM use based on the dashboard and made several suggestions. Whereas P11 would prefer a cost-benefit analysis – *"how much would that cost me an API spending versus what I pay for the ChatGPT subscription."* – P9 and P12 wish for a prompt quality analysis, with P12 stating:

> *"How precise are [...] my prompts [...]? Can I just [] get the response from ChatGPT or the response that I expect immediately? Or do I have to [] ask the same question multiple times to get [] a response back? "*

7 Discussion

Our work contributes to the growing body of HCI and education technology research by exploring how students engage in self-reflection around their interaction with LLMs. We develop PromptMirror, a personal analytics dashboard mirroring LLM use, and examine how it can support reflective practices and increase awareness of LLM use among students. Drawing inspiration from previous work in personal informatics and digital wellbeing, we present findings from a one-shot think-aloud session, where participants uploaded and reflected on their historic LLM use data via PromptMirror. We structure our investigation around two research questions: **(RQ1)** Which metrics and design elements might support students' self-reflection on LLM usage within a personal analytics dashboard? **(RQ2)** How does PromptMirror influence students' awareness of their LLM usage and the depth of their reflection?

First, we discuss the potential impact of an expanded set of metrics on reflection depth and variability of insights (RQ1). We then introduce the term *estimation gap* (RQ2) to describe students' tendency to misjudge their reliance on LLMs, before discussing our findings in relation to other technologies. Our think-aloud studies reveal a pattern of unprompted discussions about the broader impact of AI, surpassing the R2 level of reflection in visual dashboard exploration (RQ2).

7.1 Supporting Continuous Reflection (RQ1)

Currently, PromptMirror embodies rather basic metrics across the deduced five metrics categories, presented in Section 3.3. Particularly the emerged dimension *Reflection Prompts* (embodying Conversation from [12]) was left out. Despite this simplicity, the implemented metrics (presented in Figure 4) have shown sufficiency in triggering several layers of reflection, as per [32]. Moreover, participants expressed emotions ranging from curiosity, enjoyment and surprise, followed by articulating behavior change goals. As such, even low-threshold feedback bears the potential to foster reflection and surface discrepancies between perceived and actual use.

At the same time, our study raises the question of how richer or more diverse metrics, next to design decisions around format and timing, might further shape students' reflections and sustain engagement with PromptMirror over the long term. While our study participants enjoyed exploring their LLM use data, many noted that the current format lends itself more to a one-time experience than to an ongoing, dynamic resource that supports continuous reflection – both on and in action.

Addressing these points can happen on several layers. The first layer, in its closest extension of including additional (mostly visualized) metrics, such as temporal patterns of use, types of prompts entered, or correlations between LLM use and task performance, implemented straight at the source of interaction, could provide additional entry points for self-assessment. The second layer – as the expert focus group's results suggest – concerns how metrics are presented, which matters as much as which metrics are chosen, with particular emphasis on the potential of reflective prompts and guiding questions. For instance, a recent related study [42] examined the integration of LLM prompts into the learning process to help students engage in post-learning self-reflection (instead of LLM use itself, as in our work). Their participants achieved higher exam performance and self-confidence compared to control-condition assigned students, suggesting a more supportive role of LLMs for learning. Even ChatGPT itself now offers study mode [56] and break prompts [57] to facilitate learning and reflection during LLM use.

While our work within this iteration assigns PromptMirror a rather passive mirroring role (though the name PromptMirror), experts and related research envision a more active and suggestive role for such a system – one that poses questions, highlights anomalies, suggests behavior change tips or prompts peer and temporal comparisons towards more deliberate LLM use, which lead to the third layer where PromptMirror could also support not only reflection-on-action but also in-action.

7.2 Own Estimation Gap and the Third-Person Effect (RQ2)

Earlier studies on digital wellbeing revealed that users frequently underestimate the amount of time they dedicate to engaging with digital technologies [45, 53]. In the context of LLM use, recent studies suggest a similar tendency among students [22], with comparable patterns observed in our study as well.

We extend this line of research to coin the term of an *estimation gap*, that describes students' tendency to misjudge their own use of LLMs. According to our findings, the dashboard can make this estimation gap visible for some people and thereby invites reflection and potential behavior change. Nonetheless, our study did not include any repeated measures to operationalize a potential behavior change effect. Future work should therefore investigate this question in longer-term deployments that investigate and compare both perceived and actual LLM use.

Interestingly, the estimation gap became most apparent towards the end of the think-aloud session. In fact, participants that demonstrated the estimation gap expressed initial satisfaction with their LLM use patterns, while shifting their concerns over LLM *misuse* towards others. This mirrors the third-person effect [25] – where individuals misjudge their own and flag other people's behavior – notably apparent in other recent studies on LLM use in research and informal learning [47, 74]. Given the value of social comparison in self-reflective practices [11, 12], which our dashboard does not currently facilitate, we plan to include a comparative reflection feature in a future version of PromptMirror. This will enable students to compare their LLM use with anonymized peer patterns, as recommended by P3, and allow us to verify (or disprove) the persistence of the estimation gap and third-person effect.

7.3 Deeper Levels of Reflection (RQ2)

Our analysis strongly leans on the analytical concept of Visualized Self [19]. Choe et al. [19] observe that interacting with visual dashboards on personal data typically elicits reflection up to the R2 level [32]. Our think-aloud sessions yielded instances of more advanced forms of reflection (i.e., R3 and R4 from Fleck and Fitzpatrick [32]). P5 reached R4 by demonstrating critical engagement with their own practices of overuse, citing the impact of LLMs' high energy demand on the environment. Moreover, without explicit prompting from our side, several participants discussed the broader, transformative role of LLMs in their learning, aligning with the R3 level of reflection described by Fleck and Fitzpatrick [32].

One explanation for why our participants exhibited higher-order reflection might lie in the unique nature of LLMs and the narratives surrounding their use. Unlike step count or screen time, LLM use directly touches on questions of authorship [27], accomplishment [40] or academic integrity [37], introducing high stake concerns for students. Participants evaluated their own use in terms of whether they were using LLMs in the *right* way, often expressing hesitation and occasionally contrasting their responsible use with perceived misuse by others. These behaviors suggest that such reflection surpassed self-awareness towards navigating larger themes such as potential liability and repercussions. At the same time, the cultural and media discourse around LLMs (including academic discourse, which Pang et al. [58] frame as the LLMification of CHI), may amplify the described reflective behaviors. Just as public narratives about smartphone overuse shaped individuals' perceptions and self-assessment [44], the pervasive narratives around LLMs may have primed our participants to critically examine their LLM use practices – resulting in reflective behaviors that reached levels R3 and R4, exceeding those typically observed in more neutral self-tracking contexts. Finally, university guidelines for the use of AI in learning often let individual instructors decide on the policies being set [5]. In turn, the potential heterogeneity created in norms might leave students set their own innate rules about what (and what not) constitutes acceptable practice, with our study bringing students' inner dilemma to the surface.

7.4 Curated Data for Reflective Practices

Prior research in personal informatics emphasizes the inevitability of *imperfect data*. For instance, Rooksby et al. [63] describe how personal tracking systems often produce incomplete records due to technical limitations such as devices failing to log activities or software bugs. In these accounts, gaps in the record are typically unintentional, stemming from infrastructural shortcomings rather than from the user.

Our findings, on the other hand, point to a distinct pattern; rather than merely accepting imperfect data, several students in our study *intentionally curated their data traces*. This included deleting past conversations (P1, P2, P12 and to some extent P6), deleting to avoid hallucination or streamlining/diversifying their conversations. These practices resulted in gaps or distortions in the visualizations surfaced by PromptMirror, where usage timelines did not necessarily reflect the full extent of interaction. Such intentional curation raises the question of whether reflection tools should aim for completeness at all, or instead acknowledge the selective ways in which students want their LLM use to be remembered, represented, or hidden. This aligns with Elsden et al.'s notion of *Documentary Informatics* [29], which frames self-tracking not only as a means for regulation but also as a form of documentation and long-term self-expression. From this perspective, curated or partial data is not simply a flaw but a resource for narrative and remembering.

Recent work on the *secret use of LLMs* [82], demonstrates that users conceal or downplay their reliance on LLMs due to concerns about external judgment, competence, or morality. While Zhang et al. [82] describe concealment primarily in social and institutional contexts, our results suggest a complementary layer of *self-concealment*, students curate traces not only to hide LLM usage from others, but also to manage how their future selves encounter these records. This dynamic aligns with our earlier observation that students often underestimate or reframe their own reliance on LLMs. In contrast to, e.g., health tracking, where deleting past data is rare, curation appears to be a meaningful practice in LLM use, shaped by privacy concerns and anxieties around overreliance.

We argue that intentionally curated data should not be dismissed as "loss" but rather seen as a potential opportunity for supporting reflection. Designers of personal analytics for LLMs might consider ways to make curation practices visible, for instance by highlighting deleted or missing entries, or by prompting students to reflect on what motivated them to erase or fragment their traces. In this sense, "absence" itself can become a reflective artifact, pointing to moments where students grapple with tensions between transparency, self-presentation, and accountability in their LLM use.

7.5 Limitations and Future Work

Context and Generalizability. We designed and evaluated PromptMirror in an educational setting with university students. This focus allowed us to uncover the estimation gap in learning, but it also limits the generalizability of our findings to this user group. Other contexts, such as professional, creative, or everyday personal use, may equally benefit from such a dashboard and could give rise to different reflection needs and interaction patterns with LLMs. Future work could explore how PromptMirror scales beyond education and whether similar reflection mechanisms foster awareness in non-academic settings.

Objective Behavior Change Measures. Our estimation gap and the subsequent intentions to change behavior have not been followed by an objective or follow-up evidence (e.g., reduced usage or changed prompt practices). We suggest including these for future studies to examine an actionable effect of PromptMirror reflection.

Participant Population. Our study included students from Science, Technology, Engineering, and Mathematics (STEM) disciplines. The sample size and disciplinary focus limit the transferability of our insights, as learners in other domains, such as the humanities or social sciences, may rely on LLMs differently and develop distinct reflection practices. Expanding to more diverse and larger cohorts would strengthen the validity of our findings and help capture domain-specific reflection needs. Additionally, future work could investigate how younger students (e.g., middle or high school) engage in LLM reflection activities using PromptMirror, as this age group may be particularly sensitive to emerging patterns of LLM use and developing reliance.

Sentiment and Intent Analysis. The sentiment analysis model did not always perform as expected, in particular in the first iteration of our study. In several cases for these participants, automated classifications conflicted with participants' own perceptions of their conversations. For example, participants generally anticipated more positive than negative sentiment in both their own and the AI's responses. Such mismatches may have shaped their reflections in unintended ways. For the second iteration, we improved the sentiment detection method but future work could involve learners in the co-design of analytic categories to better align system outputs with subjective experiences.

Visualization Literacy and Interpretation. PromptMirror assumes that users can interpret visualizations such as temporal trends, statistical summaries, and sentiment classifications. However, differences in visualization literacy or statistical understanding may influence how participants make sense of these representations and the depth of their reflection. While most participants were able to engage with the dashboard, we did not systematically assess potential difficulties in interpreting specific visual components. Future work should investigate the effect of varying levels of visualization literacy on reflective engagement, to support diverse user groups.

Static Design Rationale. Although PromptMirror has a feature that allows users to include or exclude individual visualization types, only one participant discovered this functionality. This may have been a direct consequence of the decision not to provide active onboarding to the system. In the long term, however, we plan to actively encourage participants to include or exclude features according to their preferences. Furthermore, some participants highlighted the sensitivity of the word cloud feature — this could have been addressed with consent-based visibility in our study setting; however, we would potentially have lost the initial reaction of the participants to the visualization.

8 Conclusion

We design, develop and evaluate PromptMirror, a student-facing dashboard to mirror students' everyday use of LLMs. PromptMirror's design is informed through two focus groups with experts and students resulting in visualizing temporal, sentiment, intent, and thematic dimensions of student-LLM interaction logs.

Our key insight is that making everyday LLM use visible enables learners to recognize and reflect on their own LLM use practices, revealing an estimation gap for some: a disconnect between students' perceived and actual reliance on LLMs. However, we suggest that HCI's interest in reflecting on LLM use should go beyond only encouraging students' more responsible adoption and integration of LLMs.

In evaluating PromptMirror, we have learned and gained insights that we believe are valuable for ongoing design work in this space. We see strong potential for further innovation in reflective tools that help students *make sense of*, and *act on*, their AI use in learning settings – be it reduction or increased proficiency in the use of AI tools.

References

[1] Muhammad Abbas, Farooq Ahmed Jam, and Tariq Iqbal Khan. 2024. Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education* 21, 1 (16 Feb 2024), 10. https://doi.org/10.1186/s41239-024-00444-7
[2] June Ahn, Fabio Campos, Maria Hays, and Daniela Digiacomo. 2019. Designing in Context: Reaching Beyond Usability in Learning Analytics Dashboard Design.

Journal of Learning Analytics 6, 2 (Jul. 2019), 70–85. https://doi.org/10.18608/jla.2019.62.5

[3] Yazid Albadarin, Mohammed Saqr, Nicolas Pope, and Markku Tukiainen. 2024. A systematic literature review of empirical research on ChatGPT in education. *Discover Education* 3, 1 (26 May 2024), 60. https://doi.org/10.1007/s44217-024-00138-2

[4] Hayder Albayati. 2024. Investigating undergraduate students' perceptions and awareness of using ChatGPT as a regular assistance tool: A user acceptance perspective study. *Computers and Education: Artificial Intelligence* 6 (2024), 100203.

[5] Yunjo An, Ji Hyun Yu, and Shadarra James. 2025. Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education* 22, 1 (2025), 10.

[6] Ant Design. 2025. *Ant Design*. Ant Group. Retrieved March 29, 2026 from https://ant.design/

[7] Anthropic. 2026. *How can I export my Claude data?* Anthropic Support. Retrieved March 29, 2026 from https://support.claude.com/en/articles/9450526-how-can-i-export-my-claude-data

[8] David Armstrong, Ann Gosling, John Weinman, and Theresa Marteau. 1997. The place of inter-rater reliability in qualitative research: An empirical study. *Sociology* 31, 3 (1997), 597–606.

[9] Eric P.S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/2702123.2702234

[10] Eric P.S. Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (Vancouver, BC, Canada) *(DIS '14)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2598510.2598598

[11] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 366, 8 pages. https://doi.org/10.1145/3411764.3445673

[12] Marit Bentvelzen, Paweł W. Woźniak, Pia S.F. Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting Reflection in HCI: Four Design Resources for Technologies that Support Reflection. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 2 (March 2022), 27 pages. https://doi.org/10.1145/3517233

[13] Shreya Bhandari, Ioannis Anastasopoulos, and Zachary A. Pardos. 2025. PromptHive: Demonstrating Collaborative, Human-Centered OER Creation with LLMs. In *Proceedings of the Twelfth ACM Conference on Learning @ Scale* (Palermo, Italy) *(L@S '25)*. Association for Computing Machinery, New York, NY, USA, 392–394. https://doi.org/10.1145/3698205.3733918

[14] Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User Intent Recognition and Satisfaction with Large Language Models: A User Study with ChatGPT. arXiv:2402.02136 [cs.HC] https://arxiv.org/abs/2402.02136

[15] Stuart K Card, Jock Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann, San Francisco, California.

[16] Cecilia Ka Yuk Chan and Wenjie Hu. 2023. Students' voices on generative AI: perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education* 20, 1 (17 Jul 2023), 43. https://doi.org/10.1186/s41239-023-00411-8

[17] ChatGPT. 2025. *How do I export my ChatGPT history and data?* OpenAI Help Center. Retrieved March 29, 2026 from https://help.openai.com/en/articles/7260999-how-do-i-export-my-chatgpt-history-and-data

[18] Zixin Chen, Jiachen Wang, Meng Xia, Kento Shigyo, Dingdong Liu, Rong Zhang, and Huamin Qu. 2024. StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions. arXiv:2407.12423 [cs.HC] https://arxiv.org/abs/2407.12423

[19] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare* (Barcelona, Spain) *(PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3154862.3154881

[20] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 1143–1152. https://doi.org/10.1145/2556288.2557372

[21] Dai Clegg and Richard Barker. 1994. *Case Method Fast-Track: A Rad Approach*. Addison-Wesley Longman Publishing Co., Inc., USA.

[22] Celeste Combrinck and Nelé Loubser. 2025. Student self-reflection as a tool for managing GenAI use in large class assessment. *Discover Education* 4, 1 (26 Mar 2025), 72. https://doi.org/10.1007/s44217-025-00461-2

[23] Condens. 2025. *Condens*. Condens. Retrieved March 29, 2026 from https://condens.io/

[24] D3.js. 2025. *D3.js*. Observable. Retrieved March 29, 2026 from https://d3js.org/

[25] W. PHILLIPS DAVISON. 1983. The Third-Person Effect in Communication. *Public Opinion Quarterly* 47, 1 (01 1983), 1–15. https://doi.org/10.1086/268763 arXiv:https://academic.oup.com/poq/article-pdf/47/1/1/5382397/47-1-1.pdf

[26] Deepak Varuvel Dennison, Raycelle C. C. Garcia, Parth Sarin, Jacob Wolf, Christine Bywater, Benjamin Xie, and Victor R. Lee. 2024. From consumers to critical users: prompty, an AI literacy tool for high school students. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Vancouver, Canada, Article 2655, 9 pages. https://doi.org/10.1609/aaai.v38i21.30378

[27] Fiona Draxler, Anna Werner, Florian Lehmann, Matthias Hoppe, Albrecht Schmidt, Daniel Buschek, and Robin Welsch. 2024. The AI Ghostwriter Effect: When Users do not Perceive Ownership of AI-Generated Text but Self-Declare as Authors. *ACM Trans. Comput.-Hum. Interact.* 31, 2, Article 25 (Feb. 2024), 40 pages. https://doi.org/10.1145/3637875

[28] Erik Duval. 2011. Attention please! learning analytics for visualization and recommendation. In *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (Banff, Alberta, Canada) *(LAK '11)*. Association for Computing Machinery, New York, NY, USA, 9–17. https://doi.org/10.1145/2090116.2090118

[29] Chris Elsden, Abigail C. Durrant, David Chatting, and David S. Kirk. 2017. Designing Documentary Informatics. In *Proceedings of the 2017 Conference on Designing Interactive Systems* (Edinburgh, United Kingdom) *(DIS '17)*. Association for Computing Machinery, New York, NY, USA, 649–661. https://doi.org/10.1145/3064663.3064714

[30] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. https://doi.org/10.1145/2750858.2804250

[31] K. Anders Ericsson and Herbert A. Simon. 1998. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity* 5, 3 (1998), 178–186. https://doi.org/10.1207/s15327884mca0503_3

[32] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction* (Brisbane, Australia) *(OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 216–223. https://doi.org/10.1145/1952222.1952269

[33] Jonathan M. Golding, Anne Lippert, Jeffrey S. Neuschatz, Ilyssa Salomon, and Kelly Burke. 2025. Generative AI and College Students: Use and Perceptions. *Teaching of Psychology* 52, 3 (2025), 369–380. https://doi.org/10.1177/00986283241280350 arXiv:https://doi.org/10.1177/00986283241280350

[34] Sten Govaerts, Katrien Verbert, Erik Duval, and Abelardo Pardo. 2012. The student activity meter for awareness and self-reflection. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI EA '12)*. Association for Computing Machinery, New York, NY, USA, 869–884. https://doi.org/10.1145/2212776.2212860

[35] Neville Hatton and David Smith. 1995. Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education* 11, 1 (1995), 33–49. https://doi.org/10.1016/0742-051X(94)00012-U

[36] Yueqiao Jin, Kaixun Yang, Lixiang Yan, Vanessa Echeverria, Linxuan Zhao, Riordan Alfredo, Mikaela Milesi, Jie Xiang Fan, Xinyu Li, Dragan Gasevic, and Roberto Martinez-Maldonado. 2025. Chatting with a Learning Analytics Dashboard: The Role of Generative AI Literacy on Learner Interaction with Conventional and Scaffolding Chatbots. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*. Association for Computing Machinery, New York, NY, USA, 579–590. https://doi.org/10.1145/3706468.3706545

[37] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.

[38] Eunhye Kim, Kiroong Choe, Minju Yoo, Sadat Shams Chowdhury, and Jinwook Seo. 2025. Beyond Tools: Understanding How Heavy Users Integrate LLMs into Everyday Tasks and Decision-Making. arXiv:2502.15395 [cs.HC] https://arxiv.org/abs/2502.15395

[39] Minsun Kim, SeonGyeom Kim, Suyoun Lee, Yoosang Yoon, Junho Myung, Haneul Yoo, Hyunseung Lim, Jieun Han, Yoonsu Kim, So-Yeon Ahn, Juho Kim, Alice Oh, Hwajung Hong, and Tak Yeon Lee. 2024. Designing Prompt Analytics Dashboards to Analyze Student-ChatGPT Interactions in EFL Writing. arXiv:2405.19691 [cs.HC] https://arxiv.org/abs/2405.19691

[40] Charlotte Kobiella, Yarhy Said Flores López, Franz Waltenberger, Fiona Draxler, and Albrecht Schmidt. 2024. "If the Machine Is As Good As Me, Then What Use Am I?" – How the Use of ChatGPT Changes Young Professionals' Perception of Productivity and Accomplishment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1018, 16 pages. https://doi.org/10.1145/3613904.3641964

[41] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872 [cs.AI] https://arxiv.org/abs/2506.08872

[42] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N. Rafferty, John Stamper, and Michael Liut. 2024. Supporting Self-Reflection at Scale with Large Language Models: Insights from Randomized Field Experiments in Classrooms. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale* (Atlanta, GA, USA) *(L@S '24)*. Association for Computing Machinery, New York, NY, USA, 86–97. https://doi.org/10.1145/3657604.3662042

[43] Jérémy La Scala, Sonia Sahli, and Denis Gillet. 2025. Stimulating Brainstorming Activities with Generative AI in Higher Education. In *2025 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, London, United Kingdom, 1–10. https://doi.org/10.1109/EDUCON62633.2025.11016340

[44] Simone Lanette, Phoebe K Chua, Gillian Hayes, and Melissa Mazmanian. 2018. How Much is 'Too Much'?: The Role of a Smartphone Addiction Narrative in Individuals' Experience of Use. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 101.

[45] Heyoung Lee, Heejune Ahn, Trung Giang Nguyen, Sam-Wook Choi, and Dae Jin Kim. 2017. Comparing the self-report and measured smartphone usage of college students: a pilot study. *Psychiatry investigation* 14, 2 (2017), 198.

[46] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A stage-based model of personal informatics systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. https://doi.org/10.1145/1753326.1753409

[47] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. arXiv:2411.05025 [cs.CL] https://arxiv.org/abs/2411.05025

[48] Xi Lin. 2024. Exploring the Role of ChatGPT as a Facilitator for Motivating Self-Directed Learning Among Adult Learners. *Adult Learning* 35, 3 (2024), 156–166. https://doi.org/10.1177/10451595231184928 arXiv:https://doi.org/10.1177/10451595231184928

[49] Ulrik Lyngs, Kai Lukoff, Petr Slovak, Reuben Binns, Adam Slack, Michael Inzlicht, Max Van Kleek, and Nigel Shadbolt. 2019. Self-Control in Cyberspace: Applying Dual Systems Theory to a Review of Digital Self-Control Tools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3290605.3300361

[50] Meta Open Source. 2025. *React*. Meta. Retrieved March 29, 2026 from https://react.dev/

[51] Microsoft. 2025. *.NET*. Microsoft Corporation. Retrieved March 29, 2026 from https://dotnet.microsoft.com/

[52] Miro. 2025. *Miro*. Miro. Retrieved March 29, 2026 from https://miro.com/

[53] Catharina Muench, Johanna Link, and Astrid Carolus. 2022. The Surprise of Underestimation: Analyzing the Effects and Predictors of the Accuracy of Estimated Smartphone Use. In *Design, Operation and Evaluation of Mobile Communications: Third International Conference, MOBILE 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26 – July 1, 2022, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 171–190. https://doi.org/10.1007/978-3-031-05014-5_14

[54] Seyed Parsa Neshaei, Thiemo Wambsganss, Hind El Bouchrifi, and Tanja Käser. 2025. MindMate: Exploring the Effect of Conversational Agents on Reflective Writing. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 397, 9 pages. https://doi.org/10.1145/3706599.3720029

[55] ONNX. 2025. *ONNX*. Linux Foundation AI and Data. Retrieved March 29, 2026 from https://onnx.ai/

[56] OpenAI. 2025. *ChatGPT Study Mode*. OpenAI. Retrieved March 29, 2026 from https://openai.com/index/chatgpt-study-mode/

[57] OpenAI. 2025. *How We're Optimizing ChatGPT*. OpenAI. Retrieved March 29, 2026 from https://openai.com/index/optimizing-chatgpt/

[58] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. arXiv:2501.12557 [cs.HC] https://arxiv.org/abs/2501.12557

[59] Bernd Ploderer, Wolfgang Reitberger, Harri Oinas-Kukkonen, and Julia van Gemert-Pijnen. 2014. Social interaction and reflection for behaviour change. *Personal and Ubiquitous Computing* 18, 7 (01 Oct 2014), 1667–1676. https://doi.org/10.1007/s00779-014-0779-y

[60] Qualtrics. 2026. *Qualtrics Survey Software*. Qualtrics. Retrieved March 29, 2026 from https://www.qualtrics.com/

[61] Recharts Group. 2025. *Recharts*. Recharts Group. Retrieved March 29, 2026 from https://recharts.org/en-US

[62] Alberto Monge Roffarello and Luigi De Russis. 2023. Achieving Digital Wellbeing Through Digital Self-control Tools: A Systematic Review and Meta-analysis. *ACM Trans. Comput.-Hum. Interact.* 30, 4, Article 53 (Sept. 2023), 66 pages. https://doi.org/10.1145/3571810

[63] John Rooksby, Parvin Asadzadeh, Mattias Rost, Alistair Morrison, and Matthew Chalmers. 2016. Personal Tracking of Screen Time on Digital Devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 284–296. https://doi.org/10.1145/2858036.2858055

[64] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108

[65] Beat A. Schwendimann, María Jesús Rodríguez-Triana, Andrii Vozniuk, Luis P. Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. 2017. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies* 10, 1 (2017), 30–41. https://doi.org/10.1109/TLT.2016.2599522

[66] Donald Schön. 2017. *The Reflective Practitioner*. Basic Books, New York. https://doi.org/10.4324/9781315237473

[67] Abdulhadi Shoufan. 2023. Exploring Students' Perceptions of ChatGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access* 11 (2023), 38805–38818. https://doi.org/10.1109/ACCESS.2023.3268224

[68] Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2696–2707. https://doi.org/10.1145/3025453.3025516

[69] Ana Stojanov, Qian Liu, and Joyce Hwee Ling Koh. 2024. University students' self-reported reliance on ChatGPT for learning: A latent profile analysis. *Computers and Education: Artificial Intelligence* 6 (2024), 100243. https://doi.org/10.1016/j.caeai.2024.100243

[70] Jérémie Sublime and Ilaria Renna. 2024. Is ChatGPT Massively Used by Students Nowadays? A Survey on the Use of Large Language Models such as ChatGPT in Educational Settings. arXiv:2412.17486 [cs.CY] https://arxiv.org/abs/2412.17486

[71] Teo Susnjak, Gomathy Suganya Ramaswami, and Anuradha Mathrani. 2022. Learning analytics dashboard: a tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education* 19, 1 (14 Feb 2022), 12. https://doi.org/10.1186/s41239-021-00313-7

[72] Nada Terzimehic, Florian Bemmann, Miriam Halsner, and Sven Mayer. 2023. A Mixed-Method Exploration into the Mobile Phone Rabbit Hole. *Proc. ACM Hum.-Comput. Interact.* 7, MHCI, Article 194 (Sept. 2023), 29 pages. https://doi.org/10.1145/3604241

[73] Nađa Terzimehić, Fiona Draxler, Mariam Ahsanpour, and Albrecht Schmidt. 2023. Implicit Smartphone Use Interventions to Promote Life-Technology Balance: An App-Market Survey, Design Space and the Case of Life-Relaunched. In *Proceedings of Mensch Und Computer 2023* (Rapperswil, Switzerland) *(MuC '23)*. Association for Computing Machinery, New York, NY, USA, 237–249. https://doi.org/10.1145/3603555.3603578

[74] Nađa Terzimehić, Babette Bühler, and Enkelejda Kasneci. 2025. Conversational AI as a Catalyst for Informal Learning: An Empirical Large-Scale Study on LLM Use in Everyday Learning. arXiv:2506.11789 [cs.HC] https://arxiv.org/abs/2506.11789

[75] Alice Thudt, Bongshin Lee, Eun Kyoung Choe, and Sheelagh Carpendale. 2017. Expanding research methods for a realistic understanding of personal visualization. *IEEE computer graphics and applications* 37, 2 (2017), 12–18.

[76] Nina Valkanova, Sergi Jorda, Martin Tomitsch, and Andrew Vande Moere. 2013. Reveal-it! the impact of a social visualization projection on public awareness and discourse. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 3461–3470. https://doi.org/10.1145/2470654.2466476

[77] Katrien Verbert, Xavier Ochoa, Robin De Croon, Raphael A. Dourado, and Tinne De Laet. 2020. Learning analytics dashboards: the past, the present and the future. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (Frankfurt, Germany) *(LAK '20)*. Association for Computing Machinery, New York, NY, USA, 35–40. https://doi.org/10.1145/3375462.3375504

[78] Jason L Weber, Barbara Martinez Neda, Kitana Carbajal Juarez, Jennifer Wong–Ma, Sergio Gago–Masague, and Hadar Ziv. 2024. Beyond the Hype: Perceptions and Realities of Using Large Language Models in Computer Science Education at an R1 University. In *2024 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, Kos Island, Greece, 01–08. https://doi.org/10.1109/EDUCON60312.2024.10578596

[79] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time': Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY,

USA, 1729–1738. https://doi.org/10.1145/2858036.2858193

[80] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (April 2016), 38–44. https://doi.org/10.1145/2907069

[81] Yoctol. 2025. *react-d3-cloud*. npm, Inc. Retrieved March 29, 2026 from https://www.npmjs.com/package/react-d3-cloud

[82] Zhiping Zhang, Chenxinran Shen, Bingsheng Yao, Dakuo Wang, and Tianshi Li. 2025. Secret Use of Large Language Model (LLM). *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW163 (May 2025), 26 pages. https://doi.org/10.1145/3711061

A First Version of PromptMirror Dashboard

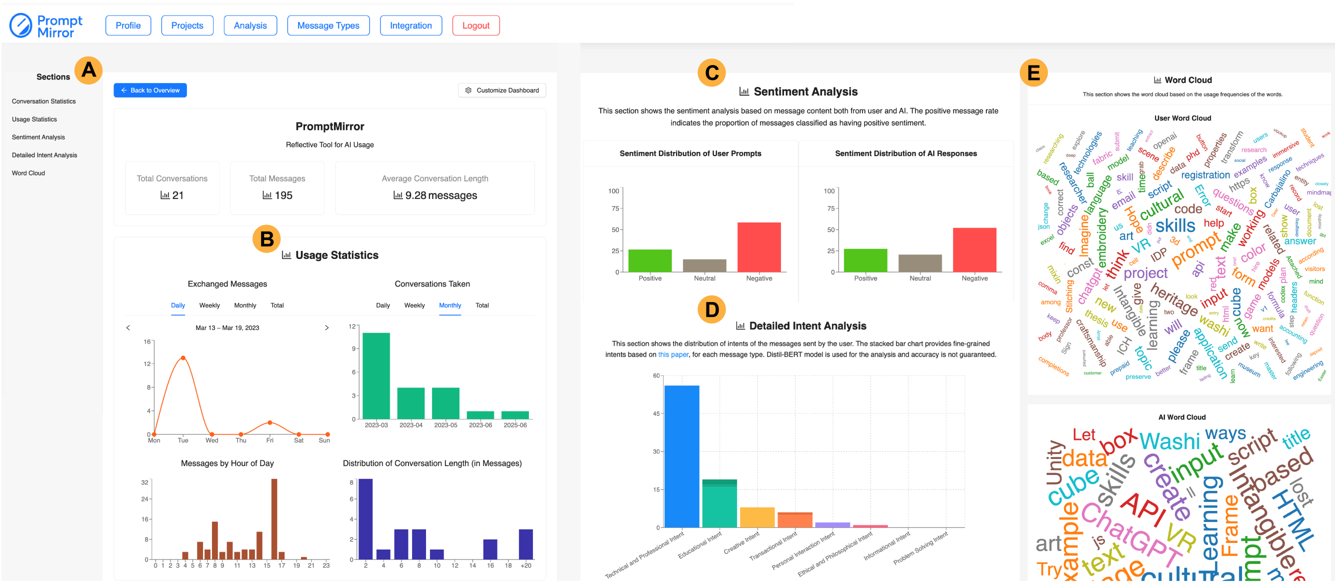


Figure 8: The *PromptMirror* dashboard visualizes students' AI usage through five reflection-oriented components: (A) Customizable navigation sidebar for accessing different analytics views; (B) Usage Statistics showing temporal patterns (daily/weekly/monthly activity, time-of-day distribution, conversation length); (C) Sentiment Analysis on user prompts and AI responses; (D) Detailed Intent Analysis classifying message purposes (e.g., technical, educational, creative...); and (E) Word Cloud visualizing prompt topics scaled by frequency.

B Tutorial for Exporting Gemini Conversation Data

- Go to Google Takeout.
- Click "Deselect all" at the top (otherwise, it will try to download your entire Google life).
- Crucial Step: Do NOT just select "Gemini" from the list. Why? Selecting only "Gemini" often exports only your saved "Gems" or settings, resulting in an empty or nearly empty file.
- Instead: Scroll down and find "Gemini Apps". If you cannot find it, find "My Activity".
- If selecting "My Activity":
  - Check the box next to "My Activity".
  - Click the button labeled "All activity data included".
  - In the pop-up, click "Deselect all" again, then search for and select "Gemini Apps". Click OK.
- Scroll to the bottom and click "Next step".
- Choose your file type (Keep it as .zip) and delivery method (usually "Send download link via email").
- Click "Create export".
- Note: It may take hours or days for Google to compile the file. You will receive an email when it is ready.