Optimizing Diary Studies Learning Outcomes With Fine-Tuned Large Language Models on the DiaryQuest Platform

Sunggyeol Oh
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0008-0033-8104

Jiacheng Zhao
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0003-5719-909X

Carson Russo
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0000-6208-958X

Michael Bolmer Jr
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0002-5681-3839

Jihoo Jeong
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0006-9789-3854

Jixiang Fan
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0008-2778-3136

Yusheng Cao
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0000-0003-0604-0946

Wei Lu Wang
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0004-7016-7794

Natalie Andrus
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0009-0008-7845-4455

Scott McCrickard
Department of Computer Science
Virginia Tech
Blacksburg, VA, USA
0000-0001-9839-7146

***Abstract*—This innovative practice full paper describes how the fine-tuned large language model (LLM) enhances student reflection and faculty efficiency in an introductory Human-Computer Interaction (HCI) course through AI-generated personalized feedback and practical data analysis. In HCI education, traditional diary studies provide a way to extract unique insights from user experiences. However, instructors, especially in large classes, often face resource limitations that impede a thorough analysis of diary entries, preventing students from receiving personalized feedback. Using a Mistral 7B model fine-tuned with more than 1,000 diary entries, the platform provides constructive feedback tailored to individual student submissions. In addition, the platform employs advanced data analytics and visualization techniques, including semantic search, thematic analysis, and semantic graphs, enabling faculty to navigate and review large volumes of student diary entries efficiently. The survey results revealed that fine-tuned LLM-generated feedback provides high-quality feedback and significantly enhances the depth of student reflections. At the same time, the AI data analytics tool improves faculty efficiency in teaching and assessment. This study highlights the potential of using LLMs to bridge the gap between a large-scale student writing dataset and human-centered learning technologies, suggesting future research on AI-based teaching environments.**

***Index Terms*—Learning technology, Computer science education, Large language models, Fine-tuning, Visualization**

I. Introduction

The diary study is a research methodology extensively employed across multiple disciplines, including computer science, psychology, and human-computer interaction (HCI) [1], [2]. This method offers an in-depth look at the qualitative aspects of user behavior by analyzing their written reflections and experiences over time. As these insights are pivotal in understanding how users interact with technology, diary studies are particularly valued in HCI [3]–[5], which help elucidate user interaction patterns with various digital interfaces and systems. Despite its proven utility, the implementation of diary studies in the context of HCI education is fraught with challenges [6]. The primary issues arise from the logistics of handling a large volume of diary entries, particularly in educational settings with large class sizes and students from diverse backgrounds [7]. These factors complicate extracting valuable insights from the entries, as educators must sift through a

significant amount of data with limited time and resources. The complexity is further exacerbated by the varying quality of the entries and the nuanced interpretations required to derive meaningful conclusions.

Learning management systems (LMS) such as Canvas [8] are often used to manage the submission and storage of these diary entries. Although these systems are effective for handling basic administrative tasks, they lack specialized features for deep content analysis, which is crucial to extracting nuanced insights from student entries. This limitation means that educators are often unable to utilize data from the diary study to their full potential, thus missing opportunities to adjust teaching methods and improve student learning outcomes based on real-time feedback and insights.

Recognizing these limitations, previous work introduced an application called DiaryQuest [9], which was designed to address some of the key challenges associated with diary studies in HCI education. DiaryQuest utilizes advanced analytical techniques, such as Latent Dirichlet Allocation [10] (LDA) and sentiment analysis [11], to automate the process of theme detection and mood assessment within student diary entries. These technologies have been instrumental in helping educators efficiently parse large volumes of text and identify prevalent themes and sentiments that reflect the student experience. However, subsequent feedback from the use of DiaryQuest revealed that while the application helps with data analysis, it does not provide immediate actionable feedback to students. The current system does not allow students to receive timely responses to their entries, which is crucial to encouraging reflective learning and improvement. Additionally, the application struggles with effectively communicating the insights extracted from the data, which limits students' ability to recognize and understand the patterns in their writing that could enhance their learning experience.

In light of these findings, this paper focused on improving DiaryQuest by integrating LLM [12]. LLMs offer considerable promise in addressing the limitations above by facilitating a more sophisticated analysis of text data. These models can generate immediate, nuanced feedback to students, highlighting areas for improvement and clarifying the key learning points of their entries. For educators, LLMs can provide a deeper, more scalable analysis of diary entries, allowing them to tailor their instructional strategies more effectively based on a comprehensive understanding of student experiences and needs. Incorporating LLMs into DiaryQuest will transform the diary study methodology in HCI education by making it more interactive and responsive. This enhancement will improve the educational experience of students by providing them with more relevant and timely feedback. However, it will empower educators to make data-driven decisions that significantly impact learning outcomes. As we continue to develop and refine this technology, it will set a new standard for the application of diary studies in educational settings, paving the way for more personalized and compelling learning experiences.

II. Background

In this section, we explore the integral role of diary studies in educational settings, particularly within the domain of HCI. Diary studies, which are categorized into feedback and elicitation types, serve as a powerful tool to capture the nuanced experiences and reflections of participants over time. This section delves into the challenges associated with traditional diary study methods, such as logistical difficulties in managing large volumes of data and inherent participation issues. Furthermore, we discuss the transformative potential of leveraging LLMs in enhancing the efficiency and depth of diary studies through the DiaryQuest platform. By integrating advanced analytical capabilities, these models facilitate more nuanced data extraction and interpretation, promising to reshape the landscape of diary-based research in educational contexts.

A. Diary Studies

Diary studies have emerged as a valuable methodological tool within the field of HCI [13]–[15]. These studies, which involve participants keeping a record of their interactions with technology over a period of time, offer deep insights into the user experience, enabling educators and students to analyze and understand the dynamic nature of human-tech relationships.

The primary significance of employing diary studies in HCI education lies in their ability to capture the temporal aspects of user experiences [7]. Unlike other research methods that offer a snapshot in time, diary studies provide a longitudinal perspective. This is particularly useful in HCI, where understanding how user attitudes and behaviors evolve with continued use of technology is crucial. For students, this method teaches the importance of context and change over time in user experiences, aspects that are often pivotal in designing user-centric systems. Furthermore, diary studies encourage a more participant-centered approach to research [16]. They allow users to report their experiences in real time and in their own words, which can lead to richer and more nuanced data than those that could be obtained through laboratory studies or structured interviews [6]. For HCI students, learning how to design and manage diary studies helps develop skills in empathy and qualitative analysis, which are the key components in user experience research. Another educational benefit of diary studies in HCI is that they promote the development of critical thinking and problem solving skills [17], [18]. As students analyze diary entries, they must consider variables such as environmental context, emotional states, and technological constraints that affect user interaction. This analysis not only deepens their understanding of human behavior about technology, but also enhances their ability to design more intuitive and accessible interfaces. Diary studies also offer a practical component in HCI education. They can be easily integrated into coursework as a means for students to conduct primary research. This hands-on experience is invaluable for students, providing them with a direct application of theoretical concepts to real-world scenarios. It prepares them

for professional roles in user experience research, where the ability to adapt and respond to user feedback is key.

However, the effectiveness of the method is occasionally hindered by several practical challenges. One persistent issue is maintaining participation throughout the duration of the study [19]. The repetitive and sometimes monotonous task of documenting routine activities can lead to decreased motivation and incomplete entries, thus reducing the reliability and richness of the collected data. Participants may also struggle with articulating deeper reflections or may revert to surface-level observations, especially without proper scaffolding or prompts. In addition, educators and researchers face significant difficulties in analyzing diary entries on scale [6], [20]. The unstructured nature and variability of qualitative data require time-intensive coding and interpretation, often lacking standardized frameworks for assessment. This makes it challenging to extract meaningful patterns efficiently or to provide timely feedback to participants. As a result, while diary studies hold strong pedagogical and methodological value, their successful implementation depends heavily on thoughtful design—balancing structured prompts with flexibility, supporting participants with guidance throughout the process, and incorporating tools or strategies that can streamline data analysis and enhance usability for both learners and facilitators. A proven method to increase participation among students is through feedback [21].

B. DiaryQuest Learning Management System

Previous research has shown that diary studies are most effective when facilitated through online platforms accessible via mobile devices, primarily due to the ease of use and the ability to integrate into participants' daily routines with minimal disruption [22]. Mobile compatibility increases the likelihood of timely and consistent entries, thereby enhancing the reliability and richness of the collected data.

One platform designed to address some of the logistical challenges associated with diary studies is **DiaryQuest** [9]. DiaryQuest streamlines the administrative and procedural aspects of diary-based research by enabling instructors to easily design, distribute, and collect diary studies. Its user-friendly interface and mobile accessibility reduce the friction for both instructors and participants, contributing to higher engagement rates and more structured data collection. However, despite its strengths in facilitating the delivery and management of diary studies, DiaryQuest currently falls short in supporting large-scale qualitative data analysis. Instructors using the platform still face significant challenges when it comes to coding, synthesizing, and interpreting large volumes of unstructured text data. The absence of built-in analytical tools, such as thematic clustering, sentiment analysis, or keyword frequency visualizations, limits their utility to drawing timely insights, especially in educational settings where instructors must evaluate reflections from multiple students within a constrained time frame.

As such, while platforms such as DiaryQuest represent meaningful progress in the logistical support of diary studies, further development is needed to bridge the gap between data collection and scalable and efficient data interpretation. Integrating automated analysis features or exporting functionalities compatible with qualitative data analysis software could significantly enhance the platform's value for instructors and researchers alike. Based on these developments, we propose to integrate AI technologies, particularly fine-tuned LLM, into the diary study platform to enhance the scalability, responsiveness, and educational value of reflective learning.

III. LLM Selection and Justification

To determine which LLM is best suited for the diary study platform, we performed a blind evaluation of six LLMs, focusing on their ability to generate feedback and summaries for diary entry submissions. These tasks, such as feedback generation, thematic analysis, and semantic search, are fundamentally focused on the ability of LLM to generate high-quality summaries and feedback that are tailored to different contexts. To generate feedback, the LLM must provide constructive and evaluative suggestions that guide individual student submissions. Thematic analysis requires the LLM to provide insight into recurring topics by summarizing each semantic cluster generated by the embedding model. Similarly, for semantic retrieval, the LLM must provide relevant search results by summarizing clusters of entries that are most relevant to the topic or keyword entered by the user. The ability of the LLM to effectively perform feedback and summarization tasks is central to the functionality of the Diary Study platform.

We evaluated and ranked six LLMs based on their performance in generating summaries and feedback for 50 diary entries. We included four open-source models (Llama 3.1 8B, Llama 3.2 11B Vision, Gemma 2 9B, and Mistral 7B) and two closed-source models (GPT-4o, GPT-4o-mini) as reference points. Each model generated a summary and one feedback for every diary entry, resulting in a total of 600 responses (300 summaries and 300 feedback) across all entries.

Six independent raters evaluated the summaries produced by each LLM for a series of diary entries. Rankings of 1 to 6 were assigned to each summary, with 1 indicating the highest quality. To aggregate these rankings into a final consensus, we employed the Borda count method, wherein each rank is assigned a score inversely proportional to its position. The scores were summed across all diary entries and assessors to compute the total Borda scores for each LLM. Additionally, Kendall's Tau coefficients were computed to measure the agreement between individual rater rankings and the aggregated ranking.

This evaluation methodology is inspired by Chatbot Arena, which comprehensively evaluates and ranks large-scale language models using pairwise comparisons, statistical aggregation, and human preferences [23]. A similar approach is effective in evaluating model performance in context-based tasks.

To ensure robustness, we performed a sensitivity analysis, recalculating the aggregated rankings individually. Furthermore, a divergence analysis identified diary entries in which

individual rating scales differed significantly from the aggregated ranking.

TABLE I
TOTAL BORDA SCORES FOR EACH LLM

LLM	Model	Total Borda Score
E	GPT-4o	1308.0
D	Mistral 7B	1127.0
F	GPT-4o-mini	1101.0
A	Llama 3.1 8B	1010.0
B	Llama 3.2 11B Vision	896.0
C	Gemma 2 9B	840.0

TABLE II
KENDALL'S TAU COEFFICIENTS FOR RATER AGREEMENT

Rater	Kendall's Tau	p-value
Rater 1	0.333	0.469
Rater 2	0.733	0.056
Rater 3	0.276	0.444
Rater 4	0.828	0.022
Rater 5	0.600	0.136
Rater 6	0.867	0.017

Rater 6 exhibited the highest agreement with the aggregated ranking (Tau = 0.867, p-value = 0.017), followed closely by Rater 4 (Tau = 0.828, p-value = 0.022). Other raters showed moderate to weak agreement.

Aggregated rankings recalculated by excluding individual raters confirmed the robustness of the results. GPT-4o (LLM E) consistently retained the top position in all scenarios. Selected rankings are presented below:

TABLE III
SENSITIVITY ANALYSIS: RANKINGS AFTER EXCLUDING INDIVIDUAL RATERS

Excluded Rater	E	D	F	A	B	C
Rater 1	1059.0	954.0	938.0	821.0	703.0	757.0
Rater 2	1066.0	921.0	883.0	863.0	778.0	719.0
Rater 6	1070.0	942.0	897.0	855.0	755.0	713.0

Diary entries with the highest divergence between individual and aggregated rankings were identified. For instance, Diary Entry 3 exhibited the greatest divergence (Standard Deviation = 2.236), followed by entries 44, 38, and 19. These entries may warrant further investigation to understand the sources of disagreement among raters.

GPT-4o (LLM E) demonstrated the highest overall performance in summarizing diary entries, supported by its leading total Borda score. Mistral 7B (LLM D) and GPT-4o-mini (LLM F) followed as strong contenders. The sensitivity analysis reinforced the reliability of the results, and the divergence analysis highlighted specific areas of inconsistency. These findings underscore the utility of the Borda count method for aggregating diverse preferences and provide a solid foundation for further comparative studies of LLM.

IV. LLM FINE-TUNING

The main goal of fine-tuning this project is to develop a lightweight LLM optimized for summarizing and generating feedback on a diary study platform. This fine-tuned LLM not only guarantees high-quality results for the summarization and feedback generation, but also balances the performances and resource efficiency by utilizing a computationally efficient model with fewer parameters compared to closed-source models such as GPT-4o and Claude 3.5 Sonnet. We chose Mistral 7B as the base model for our task, which performed best among open source models in our model evaluation.

A. Data Collection

The data set for fine-tuning came from 1,372 diary entries collected over several semesters of the introductory Human-Computer Interaction course. Following the LIMA study, which demonstrated the effectiveness of supervised fine-tuning on a small number of high-quality samples [24], we curated the 1,046 diary entries, excluding irrelevant entries, such as those containing only numeric ratings or no substantive content. This approach is consistent with the results of the AlpaGasus study, which highlights that fine-tuning "style alignment" can be achieved with small but high-quality datasets [25].

B. Data Cleaning and Preprocessing

To ensure data privacy and security, the preprocessing step involved removing sensitive information such as names, contact details, and personally identifiable information using SpaCy's EntityRecognizer component [26]. Additional preprocessing steps included standardizing formats between items and removing data from the document format that did not contribute to the analysis of diary content.

C. Fine-tuning Process

The fine-tuning process involved creating a triad of dataset structure consisting of three roles: System, User, and Assistant. The System role provides prompts for generating feedback or a summary for the given diary entries, while the User role represents the diary entries themselves. The assistant role contained the corresponding feedback or summary responses generated by LLM with over 70B parameters.

This triad of data set structure was essential for instruction, tuning the Mistral 7B model to the tasks of the diary study platform. This methodology followed knowledge distillation [27], where LLM with over 70B parameters served as the teacher model and Mistral 7B served as the student model. This process effectively distilled knowledge from a larger model to a more concise model, allowing it to maintain good responses while using fewer parameters.

We integrated Low-Rank Adaptation (LoRA) layers into Mistral 7B to preserve most of the pre-trained parameters while introducing small, low-rank updates (BA). Specifically, LoRA only modified the attention weight matrices (W_q and W_v) that are important to capture context during the self-attention process [28].

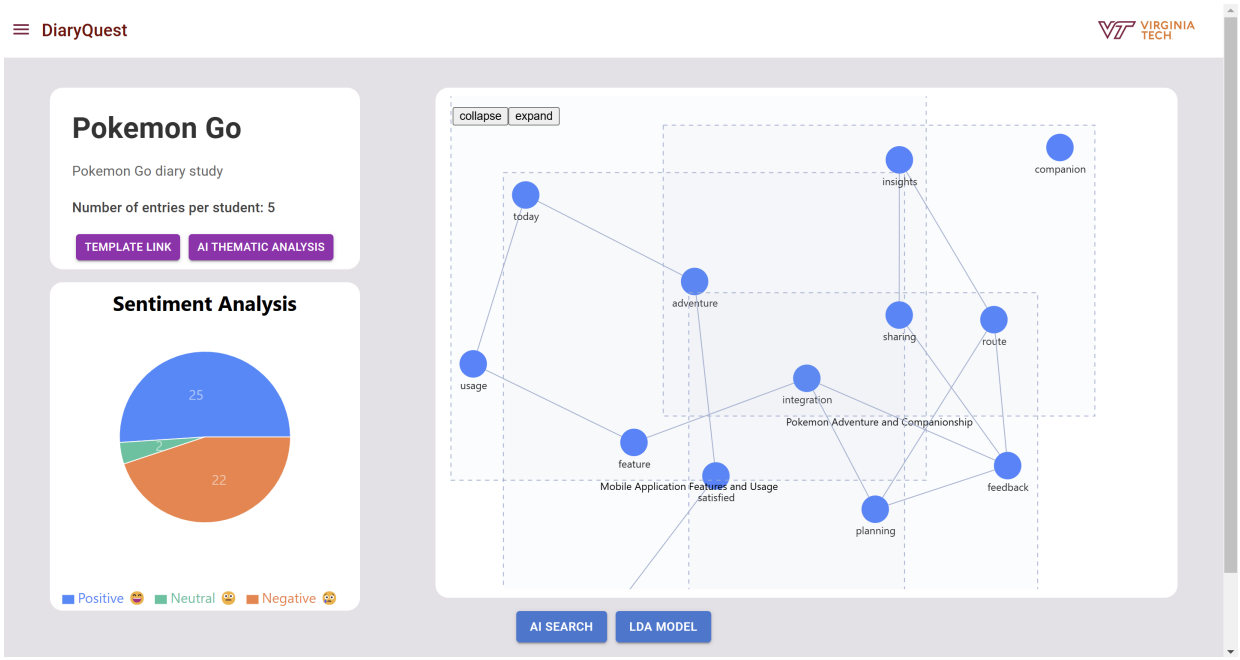

Fig. 1. DiaryQuest Analytics Dashboard

- **Parameter-Freezing:** LoRA reduces the computational load by freezing most of the weights of the Mistral 7B model during fine-tuning.
- **Low-Rank Updates:** Task-specific adaptation is parameterized by low-rank updates $\Delta W = BA$. Here, $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$, where $r \ll \min(d, k)$. This approach significantly reduces the number of learnable parameters while maintaining model performance.
- **Integration with Transformer Layers:** LoRA applies the self-attention mechanism of the transformer architecture to the key weight matrices (W_q and W_v). These layers have the greatest impact on adapting to task-specific contexts.
- **Reparametrization:** During the forward passes, the model combines the pre-trained weights W_0 and the LoRA updates ΔW: $h = W_0 x + \Delta W x = W_0 x + BAx$. [28]

The training details for the Mistral 7B model's fine-tuning process include using 13,222,350 training tokens over 10 epochs. The learning rate was set at 0.0001, which facilitated effective adaptation of the model to the specific needs of summary of diary entries and feedback generation. By fine-tuning advanced technology and tailoring it to the platform's specific goals, we have optimized the Mistral 7B model to enable efficient and effective diary entry summarization and feedback generation.

V. System Design

DiaryQuest addresses the complexity of managing and analyzing diary study data sets in education settings by leveraging machine learning techniques and natural language processing. The platform, which integrates fine-tuned LLM and visualization, improves student engagement and faculty teaching effectiveness.

The system is implemented as a unified interface with role-based access control, adapting to the needs of students and educators. With faculty access, you can view analyzes for all students using the data analysis and visualization tools, while students can only view analyses for their own submitted diaries. The data analytics dashboard provides thematic analysis, semantic search, and semantic graph visualization. On the student side, real-time AI feedback is also provided based on each diary entry submission.

A. Personalized Feedback

DiaryQuest's personalized feedback system is powered by a fine-tuned Mistral 7B language model, specifically adapted for the educational context of diary studies. By aligning the model responses with the assessment criteria for each individual diary study assignment, the platform ensures that the feedback is context-sensitive and pedagogically relevant. This approach enables instructors to provide high-quality and consistent feedback on a scale, even in large cohorts of students.

The use of a fine-tuned model, as opposed to a general-purpose LLM, offers notable advantages. Most importantly, it allows DiaryQuest to generate feedback that is tailored to the learning objectives of each assignment, ensuring alignment with course outcomes and instructor expectations. Moreover, the fine-tuning process allows the system to operate more efficiently, delivering high-quality feedback while consuming fewer computational resources, which is crucial for scalability and accessibility in educational settings.

According to Kim et al. [29], this fine-tuned feedback mechanism enhances students' learning experiences by offering actionable and constructive insights into their diary submissions. Instead of receiving generic comments, students are guided on how to deepen their reflections, clarify their thinking, or better align their entries with the study's analytical goals. This level of personalization not only supports formative assessment but also helps students improve their reflective writing and self-regulated learning over time.

By integrating AI-driven feedback with human-centered pedagogy, DiaryQuest bridges the gap between scalable technology and individualized learning, making diary studies a more sustainable and impactful method in higher education.

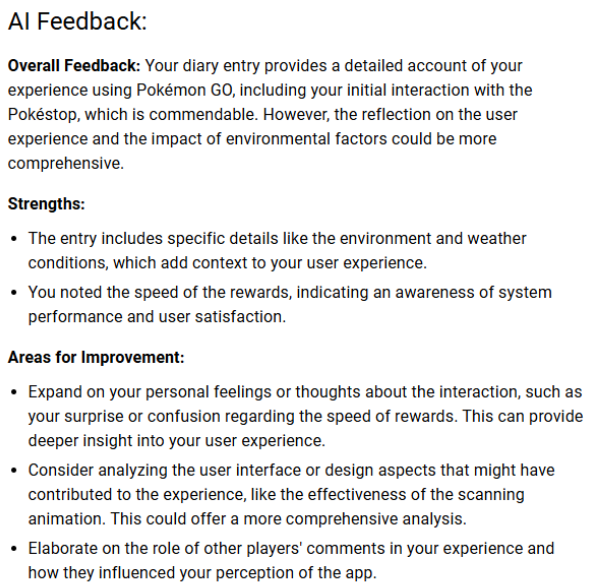

Fig. 2. Personalized Feedback Module

B. Thematic Analysis

The thematic analysis component of the platform is designed to extract meaningful themes from the entire corpus of student-submitted diary entries. This implementation represents a simplified adaptation of recent research on thematic analysis using LLMs and text embeddings [30], and is specifically tailored to operate in real time without requiring manual input or intervention from instructors.

To enable this automated analysis, the system first generates vector embeddings for each diary entry using a pre-trained language model. These embeddings capture the semantic content of the text and allow similar entries to be represented nearby within a high-dimensional space. The resulting vector representations are stored in a FAISS (Facebook AI Similarity Search) vector index [31], which enables fast and efficient similarity search and clustering.

Once embedded, the system applies the K-Means clustering algorithm [32] to group diary entries based on thematic similarity. Each cluster is then interpreted as a potential theme, providing instructors and students with an overview of dominant reflective patterns, concerns, or insights that emerge across the dataset. This process allows the platform to surface shared experiences, identify common areas of difficulty or interest, and support data-informed teaching interventions.

By automating the thematic analysis process, the platform not only reduces the analytical burden on educators but also enhances the scalability and responsiveness of diary-based learning. The real-time generation of themes ensures that both students and instructors can receive timely feedback on emerging trends within the class, further enriching the reflective learning environment.

- **Embedding Generation:** Diary entries are processed through an embedding model to produce a vector representation for each entry.
- **Vector Indexing and Clustering:** Vectors are indexed in the FAISS database and clustered using the k-means algorithm to group diary entries with similar topics.
- **Theme Identification:** A fine-tuned Mistral 7B model analyzes each clustered group of entries to derive themes for each group.
- **Interactive Exploration:** Themes are displayed as text, allowing users to hover over them to see the original diary entries that form the theme.

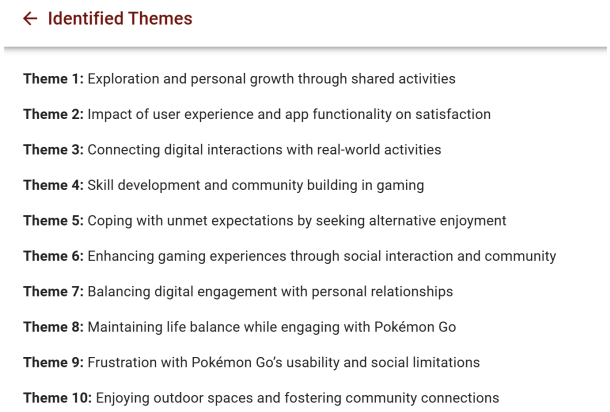

Fig. 3. AI Thematic Analysis

C. Semantic Search

Semantic search builds upon the FAISS vector space generated during the thematic analysis process to enable efficient and accurate retrieval of diary entries based on user input. When a user submits a query—whether it be a thematic phrase, a keyword, or a reflective concept—the system converts the query into a vector representation using the same embedding model used for the diary entries. It then searches the FAISS index to identify entries that are semantically similar to the query.

The retrieved diary entries are not only presented individually but are also synthesized into a concise summary using a fine-tuned Mistral 7B model. This summary helps users

quickly grasp the collective insights or patterns present in the retrieved entries, making it easier to interpret large volumes of reflective writing in context.

This feature significantly enhances the platform's usability, allowing instructors and students to perform high-accuracy searches based on abstract themes rather than relying solely on exact keywords. It supports exploratory analysis, enables rapid identification of relevant reflections, and helps surface common experiences or challenges across a group—ultimately facilitating more targeted feedback, discussion, and pedagogical intervention.

- **Query Processing:** User queries are transformed into vector representations through the embedding model.
- **Vector-Based Retrieval:** FAISS-indexed vectors are compared to identify the relevant diary entries.
- **Summarization and Citation:** The fine-tuned Mistral 7B model generates a summary of the searched items, allowing the user to see at a glance the contents of the diary entries related to the topic. Citations are provided to enable the user to check the original diary text.

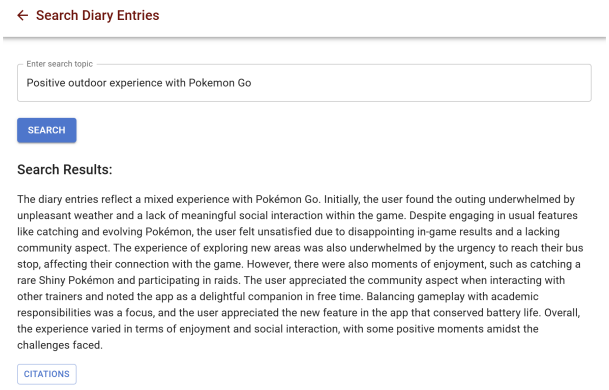

Fig. 4. Semantic Search Results

D. Semantic Graph Visualization

Semantic graph visualization intuitively illustrates the relationships between diary entries and overarching themes by applying common theme clustering to the topics extracted via LDA topic modeling. This process is carried out by an integrated LLM agent within the existing DiaryQuest system, which facilitates the automation and contextual interpretation of the extracted topics.

In this approach, each diary entry is represented as a node in the graph, and edges are drawn between entries based on shared topic distributions or thematic proximity. Clusters of nodes naturally emerge, reflecting groups of entries that revolve around similar ideas, emotions, or experiences. These clusters are further enriched by the application of LLM-assisted semantic labeling, which assigns meaningful descriptors to each theme for easier interpretation.

The resulting semantic graph provides educators with a comprehensive visual overview of how student reflections are interconnected across the data set. It enables instructors to identify dominant or emerging themes, detect isolated or underrepresented perspectives, and trace how certain topics evolve or cooccur throughout the diary submissions. By making abstract patterns in student writing visible, this tool supports data-driven pedagogical decisions and fosters a deeper understanding of the collective learning experience.

- **Topic Modeling:** LDA extracts a predefined number of topics from a diary data set [33].
- **Theme Refinement:** LLM Agent, using fine-tuned Mistral 7B, creates a common theme cluster describing topics and related diary entries and outputs it as a JSON structure.
- **Edge Computation:** Relationships between items are computed using document similarity metrics leveraging Gensim's Doc2Vec and cosine similarity algorithms [34].
- **Graph Visualization:** The G2 framework renders nodes and edges in the graph to visualize each topic and the relationship. This allows the user to click on each node to see the actual diary entries related to that topic.

VI. Evaluation

A detailed survey was conducted to assess both the quality of responses generated by the fine-tuned Mistral 7B model and the overall user experience of the DiaryQuest platform. The study involved 12 participants, comprising 2 instructors and 10 students, who had actively engaged with the platform during the course of a diary study assignment. The survey focused on two key dimensions: the relevance and usefulness of the AI-generated feedback and the usability and effectiveness of the platform interface. Participants evaluated feedback on criteria such as clarity, alignment with assignment goals, encouragement to deepen reflection, and actionable suggestions for improvement. The user experience portion of the survey examined ease of navigation, responsiveness, integration of AI features, and overall satisfaction.

To evaluate the ability of the fine-tuned model to respond effectively at varying levels of textual complexity, we classified diary entries into three length-based groups: detailed (more than 500 words), general (200 to 300 words), and brief (approximately 100 words). The participants then rated the feedback provided by the Mistral 7B model for each category using a five-point Likert scale, assessing dimensions such as relevance, clarity, depth, and usefulness.

The evaluation results revealed that the model performed well in responding to general and brief entries. In these cases, the feedback was perceived as clear, concise, and highly applicable. The participants appreciated the ability of the model to identify core ideas, provide targeted suggestions, and encourage reflective thinking without overwhelming the writer. The responses often included constructive criticism and practical guidance that supported the development of reflective writing skills and promoted personal growth.

However, the performance of the model was less robust when dealing with detailed and content-rich entries. In these

cases, the feedback tended to focus primarily on surface-level aspects such as organization, coherence, and consistency. Although these comments were still helpful, they often lacked deeper analytical insight or failed to engage with the more nuanced or layered reflections present in longer submissions. As a result, some participants felt that the feedback did not fully leverage the potential richness of their entries.

These findings suggest that while the fine-tuned Mistral 7B model is well-suited for providing feedback on shorter and mid-length reflections, further enhancements are needed to improve its handling of longer and more complex texts. Future iterations could explore the incorporation of agentic architecture or human-in-the-loop feedback mechanisms to ensure more contextually aware and depth-oriented responses for longer diary submissions.

TABLE IV
PARTICIPANT RATINGS ON DIARY FEEDBACK FROM FINE-TUNED MODEL

Quality	Avg	Std Dev	Min	Max
Detailed (500+ Words)	3.92	0.90	2	5
Ordinary (200-300 Words)	4.33	0.78	3	5
Brief (100 Words)	4.00	0.74	3	5

Participants were also asked to rate several key dimensions of the usability of the platform, including ease of use, the completeness of the features, and the intuitiveness of the user interface. The overall satisfaction score of 4.58 out of 5 reflects a strong positive response to the optimized DiaryQuest platform.

Users particularly praised the system's ability to streamline the recording and analysis of daily data through the use of intuitive charts and detailed reports. These visualization tools were highlighted as instrumental in helping users gain deeper insight into their personal activities, reflections, and behavioral patterns. Visual summaries made abstract trends more concrete, allowing both students and instructors to more easily interpret the results of diary-based studies.

In addition to improving interpretability, the visual and interactive elements of the platform were credited with contributing to a more engaging and accessible user experience. The feedback of the participants indicated that the newly introduced features, such as thematic grouping, semantic graph visualization, and automated feedback summaries, significantly improved the overall functionality and responsiveness of the platform.

These improvements collectively contributed to greater operational efficiency, allowing instructors to manage and analyze large volumes of diary entries more effectively, while helping students engage more actively in the reflective process. As a result, the improved DiaryQuest platform was seen not only as a practical administrative tool but also as a valuable pedagogical resource to support structured reflection and experiential learning.

Survey feedback highlighted the effectiveness of the AI search tool, which received an "Excellent" (5) rating for its ability to retrieve relevant diary entries efficiently. Respondents noted that the search functionality was particularly useful for narrowing down specific themes and locating diary entries with precision. The AI thematic analysis, while appreciated for its utility, received constructive feedback suggesting improvements in aligning themes more closely with the context of diary entries.

TABLE V
PARTICIPANTS RATING ON DIARY STUDY PLATFORM OVERALL

Question	Avg	Std Dev	Min	Max
Overall Functionality	4.58	0.51	4.0	5.0
UI Intuitiveness	4.17	0.72	3.0	5.0
Thematic Analysis	4.58	0.51	4.0	5.0
Semantic Search	4.33	0.65	3.0	5.0
Semantic Graph	4.00	0.85	2.0	5.0

VII. CONCLUSIONS

This study demonstrates the potential of fine-tuned LLMs to revolutionize educational methods through personalized feedback and data analytics, offering a scalable AI application framework in academia. This integration advances HCI education methodology and contributes to the broader discourse on AI in educational practices, setting a solid foundation for further explorations of LLM applications in various educational settings.

In terms of diary study methodologies, the ability of the platform to provide personalized feedback revolutionizes the traditional approach. Students benefit from LLM-generated responses that offer constructive, actionable insights tailored to their entries. This not only improves the quality of learning, but also encourages deeper reflection. For professors, the combination of AI thematic analysis and advanced search tools streamlines the grading and feedback process, reducing the manual effort of teaching assistants required to review and analyze diary entries. By making the process more efficient, the platform enables professors to focus on providing meaningful and targeted support to students.

In the context of computer science education, the platform addresses a critical challenge faced in large HCI courses: providing consistent and actionable feedback to all students. The scalability of the LLM feedback system ensures that all students receive meaningful guidance, even in courses with high enrollment. This helps student engagement and critical thinking, as students receive timely and detailed responses that encourage them to delve deeper into their reflections. By integrating AI into the educational process, the platform demonstrates how technology can enhance the learning experience and improve the quality of education.

From a broader perspective, this project offers significant insights and tools for the HCI community. The semantic graph, combined with advanced search capabilities, allows users to uncover insights and relationships that would otherwise be difficult to discern. These features not only improve the usability of diary study platforms, but also provide a framework that can be applied to other domains that require text analysis and

thematic exploration. HCI researchers can build on this work to develop more intuitive and effective tools to analyze large data sets of unstructured text.

The current evaluation involved 12 users (10 students and 2 instructors), but future versions of DiaryQuest will recruit a larger and more diverse group to further validate the platform's effectiveness and generalize the findings.

REFERENCES

[1] Y. Li, C. E. Kleshinski, K. S. Wilson, and K. Zhang, "Age differences in affective responses to inclusion experience: A daily diary study." *Personnel Psychology*, vol. 75, no. 4, pp. 805–832, Dec. 2022, publisher: Wiley-Blackwell. [Online]. Available: https://research.ebsco.com/linkprocessor/plink?id=8c7f2510-7cb8-3480-933e-3ec5f40e5c77

[2] E. W. Huff and J. Brinkley, "A Diary Study of The Teaching and Learning Experience in A High School Programming Course," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 201–205, Sep. 2021, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/1071181321651225

[3] W.-L. Wang, D. Haqq, M. Saaty, Y. Cao, J. Fan, J. V. Patel, and D. S. McCrickard, "Chatterbox opener: A game to support healthy communication and relationships," in *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2023, pp. 43–48.

[4] J. Fan, M. Saaty, D. Dunlap, and D. S. McCrickard, "Investigating an automatic assistant in computer ethics education," in *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2024, pp. 1–9.

[5] D. Haqq, J. Fan, M. Saaty, W.-L. Wang, N. Andrus, and D. S. McCrickard, "Understanding development of social games through diary studies," in *2024 IEEE Frontiers in Education Conference (FIE)*. IEEE, 2024, pp. 1–9.

[6] J. Fan, M. Saaty, and D. S. Mccrickard, "Education in HCI Outdoors: A Diary Study Approach," in *Proceedings of the 6th Annual Symposium on HCI Education*. New York NY USA: ACM, Jun. 2024, pp. 1–10. [Online]. Available: https://dl.acm.org/doi/10.1145/3658619.3658621

[7] J. Fan, D. Haqq, M. Saaty, W.-L. Wang, and S. McCrickard, "Diary study as an educational tool: An experience report from an hci course," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, 2025, pp. 311–317.

[8] Y. Cho, A. Richards, and S. Jones, "The use of canvas lms in remote academic efl writing classes: Its benefits and limitations," *Korean Journal of General Education*, vol. 17, no. 4, pp. 103–124, 2023.

[9] J. Zhao, J. Horrall, W. Gaudian, P. Jordan, P. Chavan, A. Rana, and Y. Owusu Snr, "Diaryquest: A web-based learning system utilizing diary study," in *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE TS 2025)*. Pittsburgh, PA, USA: ACM, 2025, p. 1765.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[11] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[12] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[13] L. Chen and L. Qi, "A diary study of understanding contextual information needs during leisure traveling," in *Proceedings of the third symposium on Information interaction in context*. New Brunswick New Jersey USA: ACM, Aug. 2010, pp. 265–270. [Online]. Available: https://dl.acm.org/doi/10.1145/1840784.1840823

[14] S. Carter and J. Mankoff, "When participants do the capturing: the role of media in diary studies," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Portland Oregon USA: ACM, Apr. 2005, pp. 899–908. [Online]. Available: https://dl.acm.org/doi/10.1145/1054972.1055098

[15] M. Saaty, D. Haqq, S. Dooley, J. Manalel, S. Pentakalos, S. Roshan, D. Toms, and D. S. McCrickard, "Exergames and nature." in *Nature-HCI@ CHItaly*, 2021, pp. 8–15.

[16] K. Seide, F. O. Casanova, E. Ramirez, M. McKenna, A. Cepeda, and K. M. Nowotny, "Piloting a flexible solicited diary study with marginalized latina women during the covid-19 pandemic," *International journal of qualitative methods*, vol. 22, p. 16094069231183119, 2023.

[17] H. Li, L. Yang, T. Wang, R. Xiao, L. Song, W. Xie, Z. Wang, Y. Wu, R. Su, H. Ma *et al.*, "Structured diary introspection training: A kind of critical thinking training method can enhance the pro-c creativity of interior designers," *Thinking Skills and Creativity*, vol. 52, p. 101530, 2024.

[18] L. L. Hyers, *Diary methods*. Oxford University Press, 2018.

[19] N. Bolger, A. Davis, and E. Rafaeli, "Diary methods: Capturing life as it is lived," *Annual review of psychology*, vol. 54, no. 1, pp. 579–616, 2003.

[20] A. Singh and S. Malhotra, "A researcher's guide to running diary studies," in *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction*, 2013, pp. 296–300.

[21] S. Marwan, G. Gao, S. Fisk, T. W. Price, and T. Barnes, "Adaptive Immediate Feedback Can Improve Novice Programming Engagement and Intention to Persist in Computer Science," in *Proceedings of the 2020 ACM Conference on International Computing Education Research*. Virtual Event New Zealand: ACM, Aug. 2020, pp. 194–203. [Online]. Available: https://dl.acm.org/doi/10.1145/3372782.3406264

[22] J. Brandt, N. Weiss, and S. R. Klemmer, "txt 4 l8r: lowering the burden for diary studies under mobile conditions," in *CHI '07 Extended Abstracts on Human Factors in Computing Systems*. San Jose CA USA: ACM, Apr. 2007, pp. 2303–2308. [Online]. Available: https://dl.acm.org/doi/10.1145/1240866.1240998

[23] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, and I. Stoica, "Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference," Mar. 2024, arXiv:2403.04132 [cs]. [Online]. Available: http://arxiv.org/abs/2403.04132

[24] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "LIMA: Less Is More for Alignment," May 2023, arXiv:2305.11206 [cs]. [Online]. Available: http://arxiv.org/abs/2305.11206

[25] L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, and H. Jin, "AlpaGasus: Training A Better Alpaca with Fewer Data," Feb. 2024, arXiv:2307.08701 [cs]. [Online]. Available: http://arxiv.org/abs/2307.08701

[26] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named Entity Recognition Approaches and Their Comparison for Custom NER Model," *Science & Technology Libraries*, vol. 39, no. 3, pp. 324–337, Jul. 2020. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/0194262X.2020.1759479

[27] Y. Kim and A. M. Rush, "Sequence-Level Knowledge Distillation," Sep. 2016, arXiv:1606.07947 [cs]. [Online]. Available: http://arxiv.org/abs/1606.07947

[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," Oct. 2021, arXiv:2106.09685 [cs]. [Online]. Available: http://arxiv.org/abs/2106.09685

[29] T. Kim, D. Shin, Y.-H. Kim, and H. Hong, "DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, May 2024, pp. 1–15. [Online]. Available: https://dl.acm.org/doi/10.1145/3613904.3642693

[30] I. Anakok, A. Katz, K. J. Chew, and H. Matusovich, "Leveraging Generative Text Models and Natural Language Processing to Perform Traditional Thematic Data Analysis," *International Journal of Qualitative Methods*, vol. 24, p. 16094069251338898, May 2025, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/16094069251338898

[31] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," Sep. 2024, arXiv:2401.08281 [cs]. [Online]. Available: http://arxiv.org/abs/2401.08281

[32] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, Feb. 2003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320302000602

[33] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling:

models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11042-018-6894-4

[34] M. Hanifi, H. Chibane, R. Houssin, and D. Cavallucci, "Problem formulation in inventive design using Doc2vec and Cosine Similarity as Artificial Intelligence methods and Scientific Papers," *Engineering Applications of Artificial Intelligence*, vol. 109, p. 104661, Mar. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095219762200001X