DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling

Taewan Kim
KAIST
Republic of Korea
taewan@kaist.ac.kr

Donghoon Shin
University of Washington
Seattle, WA, USA
dhoon@uw.edu

Young-Ho Kim
NAVER AI Lab
Republic of Korea
yghokim@younghokim.net

Hwajung Hong
KAIST
Republic of Korea
hwajung@kaist.ac.kr

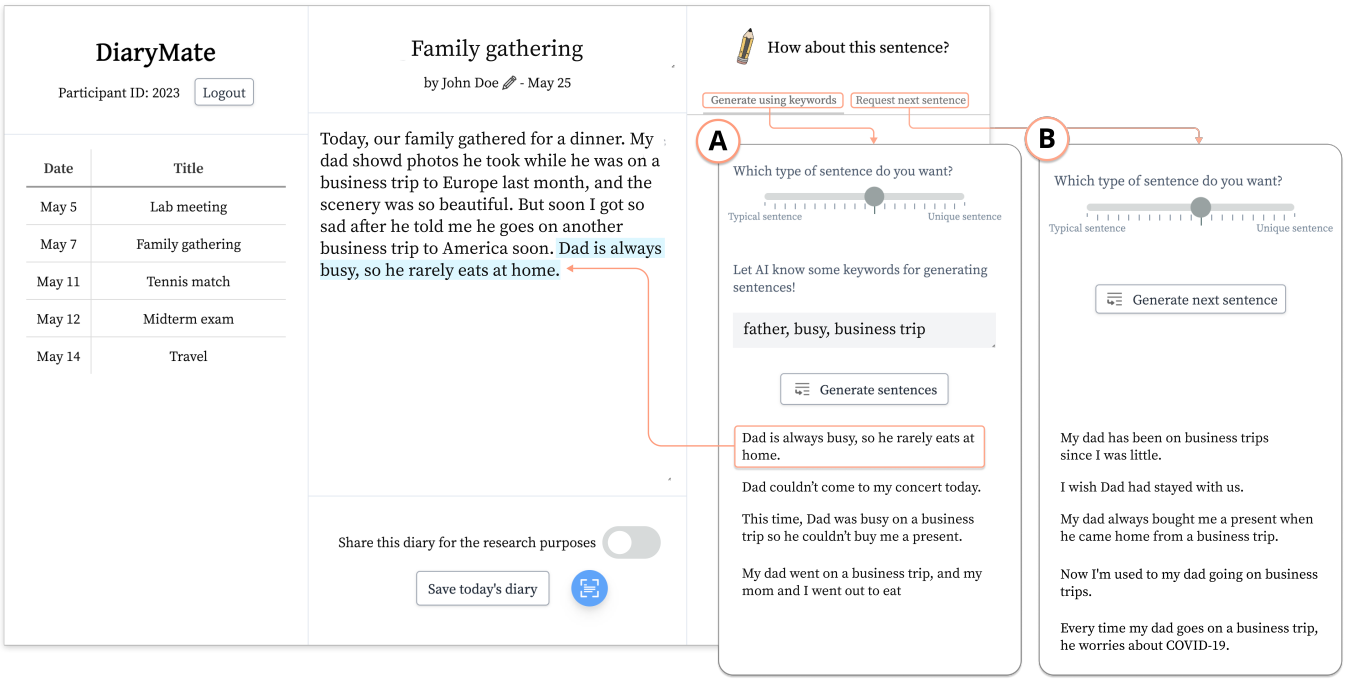

Figure 1: Key screen of DiaryMate. DiaryMate is a personal journaling assistant that leverages an LLM to assist users in writing a personal journal (A) Generate using keywords: This feature allows users to tailor sentence output using input comma-separated keywords. (B) Generate next sentences: This feature auto-generates sentences fitting the current text.

ABSTRACT

With their generative capabilities, large language models (LLMs) have transformed the role of technological writing assistants from simple editors to writing collaborators. Such a transition emphasizes the need for understanding user perception and experience, such as balancing user intent and the involvement of LLMs across various writing domains in designing writing assistants. In this study, we delve into the less explored domain of personal writing, focusing on the use of LLMs in introspective activities. Specifically, we designed DiaryMate, a system that assists users in journal writing with LLM. Through a 10-day field study (N=24), we observed that participants used the diverse sentences generated by the LLM to reflect on their past experiences from multiple perspectives. However, we also observed that they are over-relying on the LLM, often prioritizing its emotional expressions over their own. Drawing from these findings, we discuss design considerations when leveraging LLMs in a personal writing practice.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642693

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**.

KEYWORDS

Journaling, Diary, Personal writing, Human-AI collaborative writing

ACM Reference Format:
Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. 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 (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613904.3642693

1 INTRODUCTION

Pre-trained large language models (LLMs) have reshaped our understanding of how technology can interact with individuals through the natural language. These capabilities open new possibilities for supporting people in various real-world writing tasks [13]. Language models no longer merely assist writers with repetitive, simple editing tasks such as fixing typos or grammar errors; LLMs are now evolving to a point where they can actively collaborate with human writers as co-authors [13, 29, 46]. This technical advance allowed people to adopt LLMs in their writing practice in diverse domains, such as creative writing [3, 29, 46], scientific writing [17, 22], and journalism [33].

As people increasingly adopt LLMs for writing, understanding user perception and experience of writing with LLMs has become an emerging area of interest in the Human-Computer Interaction (HCI) community [8, 33, 46]. Recent work has discovered dynamics between users and LLM where users view the LLM more as an *active writer* rather than just a tool, gaining inspiration and ideas from unexpected machine-generated text [46, 47]. However, story writers hesitate to incorporate LLMs into their writing process [3]. This reluctance arises from a desire to retain control over their narrative, concerns that LLM mechanics may not align with their writing strategies [3], and perceived shortcomings in the LLM's contextual awareness [47]. These findings not only highlight the complexity and diversity of user perceptions regarding human-LLM collaboration in writing but also emphasize the need for careful implementation of LLMs as writing assistants. Specifically, the shift from viewing LLMs as a tool to perceiving them as a *companion* increases the importance of understanding user perception and experience, such as how users navigate this collaborative space, balancing their writing intent with the suggestions of the LLM.

Building on prior research, our study investigates user perceptions and experiences with LLMs in the domain of personal writing for the purpose of 'journaling,' an area that remains under-explored. Personal writing, in this context, emphasizes activities like keeping a journal, composing personal essays and letters, blogging, and penning autobiographies. People chronicle their personal histories and seek to make sense of their lives and their place in the world. These forms are inherently introspective, personal, subjective, and emotionally rich [4, 38]. Therefore, in personal writing, users may react differently to the collaboration with LLMs, given the deeply personal nature of the content. For instance, in journal writing, where grasping the writer's context is crucial, an LLM might generate content that doesn't perfectly resonate with the writer's specific situation. This arises because LLMs base their suggestions on statistical projections from vast datasets [27, 35]. However, there remains a gap in understanding people's perceptions and experiences with LLMs in the context of personal writing. Therefore, we centered our work on understanding how people perceive and utilize the outputs of LLMs in personal and retrospective writing tasks.

To this end, we developed DiaryMate, a personal journaling assistant that leverages an LLM to assist users in writing a personal journal. DiaryMate was designed as a technology probe (TP), for the purpose of understanding the perception and experience of writing with LLM in a real-world setting [24]. Through probe deployment and follow-up interviews, we collected data on how 24 participants used DiaryMate for ten days and their perceptions and experiences. We then discuss the design considerations about applying an LLMs in the journal writing context based on findings.

The major contributions of our study are as follows:

- DiaryMate, a web-based application where users can write a daily journal with assistance from an LLM
- Results from the field deployment study conducted for ten days with 24 participants using the TP approach, where we gathered qualitative and quantitative data on how people embrace and leverage LLMs for journal writing
- Future opportunities and design considerations of using an LLM for personal journal writing based on participants' experiences using DiaryMate

Through this study, we offer insights into how individuals interact with, perceive, and derive value from LLM-assisted writing tools in a personal context. This exploration extends the current understanding of LLM applications beyond professional or creative writing domains to include more personal and subjective writing context, an area that has been less explored in the HCI research field. Furthermore, by focusing on personal journal writing, the study contributes to a broader discourse on the role of LLMs in supporting introspective and reflective practices.

2 RELATED WORK

2.1 Personal Journal Writing

Journal writing is a classic method for individuals to record not only their observations, travels, and findings, but also their overall daily experiences and thoughts [38]. By disclosing their own experiences and inner emotions through the journal, people can benefit by improving their self-expression in a secure environment, which ultimately enables them to explore meaningful insights by revisiting past experiences [44, 45]. Moreover, the process of journaling aids in pinpointing personal challenges and serves as a therapeutic outlet for stress relief [1, 42].

Then, what are the characteristics of the journal in terms of writing style? Prior research generally offers: writing regularly for a certain period of time, writing down your deep feelings and thoughts as openly as possible, and writing without worrying about writing skills such as grammar and spelling and evaluation [20, 37, 44]. As such, journaling gives writers a lot of freedom and seems easy to do, but it can be overwhelming and challenging for some writers who are not accustomed to writing. Empirically, people report difficulties in journaling with starting the first sentence, with the ability to write and express themselves, and with staying

motivated [44]. Furthermore, understanding and expressing one's inner self can be complex for some people, as individuals differ in their ability to identify, express, and organize their emotions and feelings [31, 39]. Therefore, processes like journaling, which involve reflecting on past experiences and finding meaning within them, are not necessarily easy or natural for everyone [28].

In the field of human-computer interaction (HCI), studies have been presented on the use of technology to support people's expression of past emotions and experiences by writing. For example, it has been reported that creating a social atmosphere using conversational agent technology can play a positive role in motivating an individual's self-exploration and expressive writing [36]. In addition, social support received from peers in the online community also positively affects user engagement in journaling [30]. An interaction design was also proposed to mediate the negative impact of recalling past adverse situations/memories while journaling using sound generation from data analysis, which improved the enjoyment and pleasure of the writing experience [19]. These studies comprehensively show how technology can help people's journaling process.

Building on the foundation laid by prior work, our research primarily focuses on the potential of LLMs to assist human writers in personal journal writing. Our aim is to delve into how LLMs' text-generation capabilities can aid users in articulating complex emotions or thoughts that are otherwise hard to express in writing. Furthermore, we aim to investigate the extent to which LLMs, with their ability to comprehend the context of existing writing, can contribute to the enrichment of personal journals by generating coherent and contextually relevant sentences. While LLMs have been shown to enhance general writing abilities—by promoting varied vocabulary use and clarifying ambiguous ideas [10, 17, 29]—their application in the realm of personal journaling remains underexplored. Furthermore, journal writing is deeply personal and can be influenced by one's environment, so it is crucial to consider how LLMs are used in this context thoughtfully. To understand how users incorporate LLMs into their daily journaling and gauge their real-world benefits, we utilized the technology probe (TP) approach [24], which provides insights from actual user interactions.

2.2 Language Models for Writing Support

Advances in natural language processing (NLP) have enabled a range of technological writing aids, evolving from basic spell-checking to sophisticated style and grammar checks [11–13]. With the emergence of LLMs, these tools, once limited to error detection, now offer text rewriting and generation capabilities [13].

One emerging research topic in HCI is to explore how LLMs can be used for collaborative writing across various writing domains. This includes creative fields like story writing [10, 29, 47] and metaphor writing [16], and fiction [46], as well as non-fiction areas such as argumentative essays [29] and scientific writings [17]. Early studies first studied how writers work with LLMs, suggesting interaction methods for human-LLM collaborative writing like infilling, continuation, and elaboration. These interaction methods are now commonly used when developing human-AI collaborative writing systems [47]. Later research expanded its focus to encompass the overall writing process and diverse contexts. For instance, Biermann [3] proposed a design for an AI companion for story writing. In this work, they presented a human-AI collaborative writing support approach that reflects domain-specific characteristics of story writing, such as defining scenes or character relationships. In other studies, LLM-based writing assistance has been integrated and contextualized within social media and online communities to promote empathetic communication between users [41].

These studies highlight the importance of understanding how best to tailor and apply LLMs across diverse writing domains and contexts rather than merely employing them for general writing support. Additionally, the use of LLMs is branching out into diverse realms of writing, such as social media [41], where personal resonance is crucial. In this study, we delve into the unexplored potential of LLMs in personal writing, examining both its opportunities and challenges. Through this study, we aim to offer insights into designing interactions for human-AI collaborative writing, particularly in personal and reflective writing domains.

2.3 User Perception and Experience while Writing with LLM

With the role of writing assistance transitioning from a tool to a collaborator or even a coauthor, understanding how human writers perceive and accept LLMs in the writing context has become a critical research agenda. In the realm of creative writing, writers utilize the outputs of LLM writing assistance as sources of inspiration, incorporating them as new characters and details or as avenues for new story developments [8, 29, 47]. However, limitations in the control and nuances of LLM's sentence generation can sometimes hinder writers from reflecting on their writing strategies, causing some to be hesitant in adopting the technology [3]. Consequently, recent HCI research emphasizes the importance of acknowledging the writer's intention and control by designing more controllable interactions that reflect the writer's intent in the LLM's output [10]. In non-fiction writing areas like science writing and argumentative writing, LLM outputs help crystallize the writer's thoughts and offer fresh perspectives [17]. Yet, collaboration with LLMs in areas where factual accuracy is crucial brings concerns of plagiarism and hallucination [15, 40]. This also highlights issues related to determining the original authorship of LLM outputs and their usage [43]. As complex user perceptions related to writing are reported, how can the quality of human-AI collaboration in writing be determined and evaluated? And what evaluation criteria are needed? Such critical questions have been raised by researchers [47]. Stemming from this need, recent LLM research has presented datasets of human-LLM interactions to understand the LLM's sentence generation capabilities and subjective interpretations in a more systematic and data-driven approach [29]. To summarize, firstly, the LLM is perceived not merely as a tool but as a companion, and secondly, the collaboration and perception between human writers and the LLM can vary widely based on the writing domain and the writer's intention and perspective.

While most LLM research on user perception and experience has focused on areas like creative writing, our study aims to explore the less-examined area of personal writing, specifically journal writing. Journal writing differs from others, being deeply personal

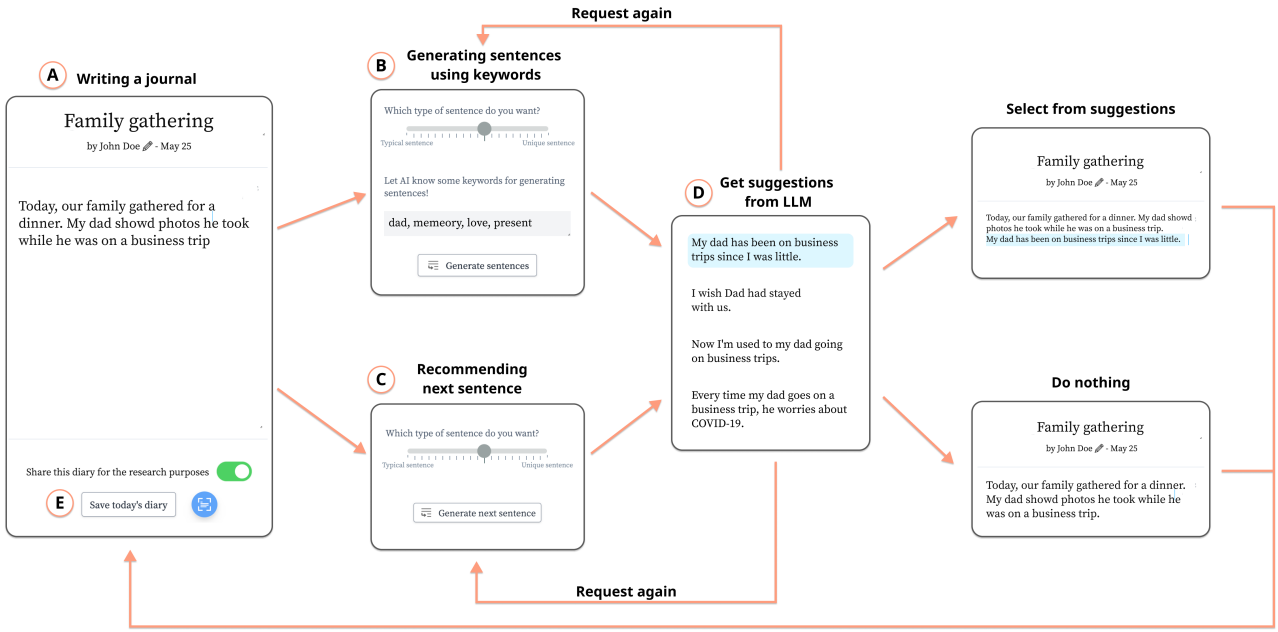

Figure 2: Diagram of DiaryMate showing how users write entries and use features (*i.e.*'Keyword-based sentence generation' and 'Suggested next sentences'). Users can either incorporate the text recommended by LLM into their diary, ignore it (just for reference), or request a new one.

and emotionally charged [38]. From the perspective of human writers, journaling involves the ability to delve into and reconstruct meaningful and reflective narratives from one's past and present thoughts and daily experiences. Therefore, the writer potentially requires or expects a different kind of assistance compared to the other writing domains. Our study aims to understand how people accept and perceive LLMs when writing in a reflective and personal context like journal writing.

3 DESIGN OF DIARYMATE

3.1 Technology Probe (TP)

Our system, DiaryMate, was designed as a prototype to probe how a newly introduced technology (*i.e.*, LLM) can be perceived and used in daily journal writing. On such an account, we determined high-level TP design guidelines drawing on prior work [24, 26]. We aimed to design a system that (1) includes a core functionality of the LLM so that users can concentrate on the central concept of the LLM, (2) allows participants to engage in open-ended and exploratory use to help researchers examine how users construct the meaning and value of LLM in journal writing, and (3) collects various types of user-generated data including participants' diary entries and interaction logs to understand meaning-making processes better. The following sections describe DiaryMate's features and implementation details.

3.2 Overall Concept and Interaction Flow

DiaryMate is a writing assistant that supports users' journal-writing processes. Using DiaryMate, users can get AI assistance while writing their daily diaries. We intend to use DiaryMate to explore how people embrace and perceive recent LLM technologies in journal writing.

To inform the design of an AI-assisted writing feature, we first clustered the types of features of existing LLM-based co-authoring systems [10, 16, 17, 29, 46, 47], as well as the capabilities of widely used LLMs [5, 6]. Then, we discussed the criteria for clustering until we reached an agreement among four researchers. Through this process, we were able to elicit two major approaches to implementing the system: *User-guided authoring* and *Unguided authoring*. User-guided authoring lets users operationalize factors in the system, so that they can intentionally guide the output. For example, in the TaleBrush system [10], the user's intent can be included in the LLM's sentence generation through the drawing of fortune lines. Likewise, we decided to simplify the process for users to establish their intentions by asking users to add keywords that they wanted. On the other hand, unguided authoring simply generates the next sentences based exclusively on the existing contents as proposed in [29, 47].

Based on the design goals suggesting both guided and unguided authoring capabilities, we designed and instantiated DiaryMate's writing flow such that users can undertake the following procedure. First, just as in conventional diary writing, users can enter a title and freely add diary content in our interface (See Figure 2-(A)). In this process, whenever users need assistance from the AI, they

can send a sentence generation request by several pre-defined keywords (See Figure 2-(B)) or without specific inputs except texts they have already generated in the text field (See Figure 2-(C)). Once completed, users can save the data (See Figure 2-(E)), and the system provides feedback (See Figure 1-(C)), a summarizing and empathizing message generated based on the diary content that the user saved.

3.2.1 Generating sentences using keywords. In the "Generate using keywords" tab menu, users are shown a slider to manipulate temperature and keywords for sentence generation (See Figure 2-(B)). Using the slider, the user can change the temperature to any level between "typical sentence" (low temperature) and "unique sentence" (high temperature). The system sources keywords from users, which later serve as material for generating sentences. Although there is no strict requirement for keywords, users are encouraged to separate keywords by a comma (*i.e.*, chores, tired, roommate, ...), notified with placeholder text, to make the structure coherent with the format of our trained texts.

Next, the user clicks "Generate sentences," which subsequently returns up to five sentence recommendations (See Figure 2-(D)). If the user is satisfied with any of the recommendations and wants to add them to the content text field, they can click the sentence; otherwise, users can start the process over or skip the recommendations.

We intended this approach to allow the system to effectively use keywords as a writing aid without limiting the model's ability to produce original and meaningful text relevant to the user. By providing the model with a seed of relevant keywords, we were able to guide its output in a way that was contextually relevant and coherent while still allowing for creativity and flexibility in the text it generated.

3.2.2 Recommending next sentences. This function automatically generates relevant sentences based on the existing text. Users can simply click "Generate next sentences" to generate the next sentences corresponding to the existing text (See Figure 2-(C)). Similar to keyword-based sentence generation, users are asked to adjust the temperature slider to the desired level. Once clicked, users are given five sentence candidates, analogous to that of keyword-based recommendation. The user can choose one among up to five sentences recommended by DiaryMate, which is appended to the content upon click.

We intended this feature to augment the introspective journaling process by offering sentence suggestions. When a user encounters a moment of hesitation or seeks to expand their thought process, they can utilize this feature. It leverages the context of the existing text to generate sentences that are coherent with the user's narrative and emotional tone. Especially, rather than just recommending a single sentence and prompting its use, we designed the Recommending next sentences feature as a tool that allows users to explore a variety of sentences they could use in their journals.

3.2.3 Receiving feedback about a completed entry. Users can complete their journal entries by actively collaborating with AI-assisted sentence-generation functionalities (*i.e.*, Generating sentences using keywords, Recommending next sentences). Once complete, users can record their diaries by clicking the "Save today's diary" button. When a diary is saved, a pop-up screen notifies the user that their diary has been saved successfully. Here, DiaryMate also briefly summarizes based on their diary content.

3.2.4 Model selection & prompt setup. Our system is targeted to users in South Korea, whose main language is Korean. As such, we decided to utilize a large language model (LLM) that can be used in the Korean context and is comparable to similar alternatives (*e.g.*, GPT-3 [6]), so as to ensure the generalizability of our study. As such, we decided to use HyperCLOVA [27], an LLM deployed in the Korean language. Consisting of 82B parameters, HyperCLOVA is trained on 560B Korean tokens to accommodate diverse Korean few-shot learning tasks. In addition to metrics that are comparable to existing, widely used LLMs, the operationalizable environment (*e.g.*, user-configurable parameters, temperature, max-tokens, and penalties) is analogous to its alternatives. Thus, we believed that such similarities in configurations would make our study more generalizable to LLMs targeted to other languages.

3.3 System Implementation

DiaryMate was developed as a web application. The system is built upon Svelte (Javascript-based framework), and the LLM (HyperCLOVA) is connected and implemented on the Python web server. Once the user requests sentence recommendations or completes a diary, the web app sends a request to the web server and returns the results, and the data is then populated in the web app. All interaction logs and diary data are saved onto the Google Firebase database.

4 METHOD

4.1 Field Deployment Study

We conducted a 10-day field deployment study using DiaryMate to explore the opportunities and challenges of using an LLM in the journal writing context. The purpose of our field study was not to evaluate the system's usability, but to use DiaryMate as a TP to understand how people use technology in their journaling practice and what perceptions and desires they had toward an LLM. To this end, we provided the participants with a minimum guide and requirements, and allowed them to use the system as freely as possible.

The field study proceeded according to the following steps: (1) A 30-minute introductory session introducing the DiaryMate system, (2) the process of using DiaryMate in daily life for ten days, and (3) a 40-minute follow-up interview to understand the influence of an LLM on journal writing experience and the perceptions of the LLM. The introductory session was held in an offline space with no more than five participants. Interview sessions were conducted using online ZOOM meetings. The IRB of the researcher's university approved the procedure of our study, and informed consent was obtained from all participants. We will further elaborate on the detailed measures in the Ethical Considerations section.

4.2 Collected Data

4.2.1 Application usage log. We focused on understanding how participants interacted with LLM in DiaryMate and how it affected their writing. As such, we embedded a tracking code in the system

to measure user behavior when using DiaryMate. More specifically, we collected the following event logs from the system: (1) interaction with LLM, including input data (*i.e.*, keyword, pre-existing text) when a user requests an LLM feature, temperature parameter settings, a list of sentences suggested by the LLM, and the sentence selected by the user, (2) diary data, including the title and contents of the diary generated by the user and feedback messages generated by the LLM, and (3) data on overall user engagement, such as log-in logs and diary review logs.

4.2.2 In-situ experience log. In addition to the log data automatically collected by the system, we added an in-situ experience logging function (The blue capture button is located in the bottom area of the center column. See figure 1) that can collect qualitative user experiences feedback on the fly. During the orientation, participants were guided on how to capture moments of interaction with DiaryMate. We informed them that, if they wished to document moments that were notably impactful, irrespective of whether these experiences were positively or negatively connoted, they could press the capture button to log these events; upon pressing the button, a popup window appears where there is a checkbox question asking participants to select which feature prompted them to press the capture button (options include Generating sentences using keywords, Recommending next sentences, the overall process, or reviewing the diary). Following this, two distinct text fields are offered. The first field is designed to document positive experiences with artificial intelligence, while the second field is for recording any negative experiences. Users are allowed to freely write in whichever field they prefer, based on their experiences.

4.2.3 Follow-up interview. We conducted semi-structured interviews with participants after ten days of using DiaryMate to understand how they used the system in their daily lives. The interviews were conducted using Zoom video meetings that lasted approximately 40-50 minutes. We structured our interview protocol around the following four themes: (1) participants' existing journaling habits, (2) overall impression and user experience of DiaryMate, (3) the impact of LLM on the diary writing process and how participants used LLM, and (4) expectations, wishes, and future ideas for using LLM for journal writing.

4.3 Participants

We posted recruitment announcements in a university's online communities and bulletin boards. Participants were recruited from people who met the following criteria: (1) undergraduate/graduate students over 20 years of age (2) who are able to access DiaryMate in a desktop or laptop environment during the study period. We conducted an introductory session with 26 participants; two dropped out during the field deployment study. 24 participants completed the experiment to the end. We ran follow-up interviews for all willing participants; 20 participated in the interview session. We provided compensation of 25 USD to all participants who participated in the field deployment study for ten days and added 10 USD compensation to the interview participants.

The participants were 14 females and 10 males. Their ages ranged from 20 to 31 years ($M = 23.25$; $SD = 3.01$). Fourteen were undergraduate students, seven were masters students, and three were doctoral students. All participants were majoring in fields related to Science and Technology and had basic knowledge of computing. More specifically, nine had completed AI courses above the undergraduate level, and seven had experience in AI-related projects. They had varying degrees of journal writing experiences. In the preliminary survey conducted before the experiment, six (25%) participants had never written a journal, nine (37.5%) participants occasionally (more than once a month) wrote a journal, six (25%) participants had kept a journal more than thrice a week, and three (12.5%) participants answered that they had kept a journal more than five times a week.

4.4 Analysis

We conducted quantitative and qualitative analyses to understand how participants used DiaryMate in their daily lives and how it affected their journaling experience. We first conducted a descriptive statistical analysis to understand the usage patterns and to obtain an overview of how participants used the system during the 10-day field deployment study. We also used t-test and Pearson correlation analysis to compare each LLM feature's characteristics and usage patterns (*i.e.*, Generating sentences using keywords, Recommending next sentences). Statistical analysis was performed using GraphPad Prism version 9.0[1] and R studio. Qualitative analysis was conducted from two perspectives. First, to deeply understand participants' subjective evaluations, perceptions, and experiences of using LLM, we analyzed the combination of qualitative data from interviews and in-situ experience logs. Second, to examine the characteristics of sentences that participants brought in their diaries among the LLM generated, we analyzed log data about the sentences that participants brought those sentences into their journals. For qualitative analysis, our study employed thematic analysis to investigate the use and perceptions of LLM in journal writing. For that, three researchers read the transcript and generated initial codes using ATLAS.ti Mac (Version 22.0.6.0) [2]. The whole research team then had a discussion to resolve any disagreements between initially generated codes and finalized codes. Themes then were generated based on these open codes. During this process, we could identify statements that revealed the use cases of LLM in journal writing practice and users' desire and perception of the LLM. We structured our themes around understanding (1) how participants use an LLM in journaling contexts and (2) how participants want LLM to behave in journal writing.

4.5 Ethical Considerations

We acknowledge that this study addresses ethical concerns, even though the institutional review board approved it. In conducting this research, we have taken care to address the potential ethical concerns. These concerns include the potential for the LLM to generate offensive or violent content [9, 14]. To address this risk, we informed participants in the introductory session about the potential for the LLM to behave unexpectedly and provided instructions for reporting any problematic words that may occur. Our protocol, which the IRB approved, includes detailed measures for monitoring participants and a follow-up procedure that incorporates the

[1] https://www.graphpad.com/scientific-software/prism/
[2] https://atlasti.com/

Table 1: Summary of our study participants. The AI Proficiency level of participants is indicated by whether they have taken AI-related courses above the undergraduate level and/or participated in AI-related projects. Journal writing frequency is classified as Occasionally: More than twice a month, Regular: 3 or more days per week, Daily: 5 or more days per week.

PID	Age/gender	Education level	Proficiency in AI	Journaling frequency
1	23/F	Undergraduate	-	Daily
2	24/F	Masters	Attended AI-related courses	Regularly
3	21/M	Doctoral	Attended AI-related courses	Occasionally
4	21/M	Masters	-	Occasionally
5	21/F	Masters	Attended AI-related courses & Participated in AI-related projects	Occasionally
6	31/M	Masters	Attended AI-related courses	Occasionally
7	23/M	Undergraduate	-	Never
8	26/F	Undergraduate	-	Never
9	22/M	Undergraduate	-	Regularly
10	31/F	Undergraduate	Attended AI-related courses	Occasionally
11	25/F	Undergraduate	Participated in AI-related projects	Regularly
12	25/F	Undergraduate	Participated in AI-related projects	Never
13	23/F	Undergraduate	-	Daily
14	23/F	Masters	Participated in AI-related projects	Occasionally
15	22/F	Doctoral	Participated in AI-related projects	Regularly
16	23/F	Undergraduate	-	Occasionally
17	23/M	Undergraduate	Attended AI-related courses	Never
18	21/M	Masters	-	Regularly
19	18/M	Doctoral	-	Occasionally
20	26/F	Undergraduate	-	Occasionally
21	23/M	Undergraduate	Attended AI-related courses	Daily
22	21/M	Masters	Attended AI-related courses & Participated in AI-related projects	Never
23	21/F	Undergraduate	Attended AI-related courses	Regularly
24	21/F	Undergraduate	Participated in AI-related projects	Never

university's mental health care facility. The protocol ensures that the research supervisor and co-researchers continuously observe the interaction between the participant and the researcher. If any mention of self-harm, suicide, or harm to others is found, we have prepared further measures that include full payment of the participation fee and termination of the experiment, as well as follow-up actions through collaboration with the campus mental health center. Furthermore, we used an AI filter provided by HyperCLOVA to detect and block potentially sensitive or harmful content from being generated by the LLM. This filter was trained on a large dataset of offensive and harmful language and was able to effectively prevent these types of content from appearing in our study. Lastly, in recognition of the potential for journals to contain personal information, we have put measures to ensure that participants' privacy is protected. Whenever users saved their diary content, we asked if they would share it for research purposes. Only the data that users consented to share was analyzed. Any data not consented to be transferred was saved in a separate location on our server, to be accessed only when the users wanted to see their entries again.

5 RESULT

In the results section, we first summarize the descriptive statistics to illustrate how participants engaged in DiaryMate. Then, we report how participants perceived LLM in the journaling process and their wishes and expectations about the role of LLM in the journal writing context. Each participant is referred to by a participant number, followed by their age and gender in parentheses. For instance, 'P01, 30(M)' denotes participant number 01, who is a 30-year-old male.

5.1 Descriptive Summary of DiaryMate Usage

Through the field deployment study, 24 participants wrote 215 journals during the ten days of the study. The average length of the

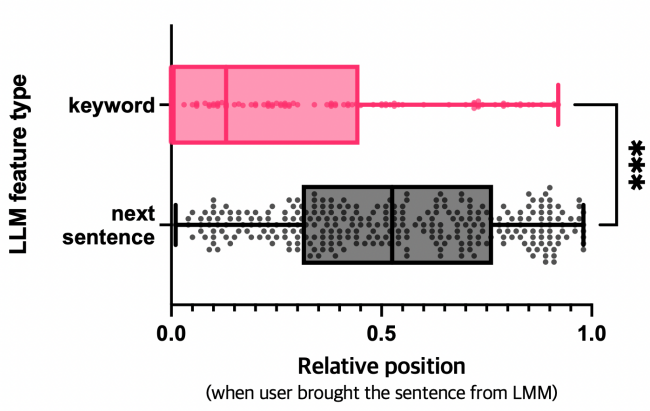

Figure 3: Box whisker plot illustrated when participants used each LLM feature (*i.e.*, Generating sentences using keywords, Recommending next sentences) during the writing of a journal. The x-axis represents the start and end positions of the journal, with 0 representing the start and 1 representing the end. Participants used sentences generated from the Generating sentences using keywords in a significantly earlier phase of writing than sentences generated from the Recommending next sentences feature. Statistically significant results are reported as *: $p < .05$, **: $p < .01$, ***: $p < .001$

journal was 426.62 syllable count [3] (SD = 320.00, min = 59, max = 3323). On average, each participant wrote 8.96 diaries during the experiment (SD = 1.52, min = 5 [P12], max = 12 [P16, P24]). Nine participants (37.5%), including seven female participants, wrote more than ten journals (at least one journal per day), 13 participants (54.2%), including six female participants, wrote more than eight journals (average of more than 0.8 diaries per day), and two (8.3%), including one female participant, wrote less than seven journals (average of 0.7 or less per day). Overall, the participants were highly engaged in daily journaling tasks during the 10-day field study. In the following section, we summarize the findings of the statistical analysis conducted on the log data.

5.1.1 When were the LM features used? During the study, our participants requested a total of 932 LLM features (*i.e.*, Generating sentences using keywords: 329, 34.8%; Recommending next sentences: 603, 64.8%). Among them, there were 434 cases (Generating sentences using keywords: 140 cases, Recommending next sentences: 294 cases) in which the participant selected a sentence out of a total of 932 LLM requests. We then further analyzed 434 LLM usage cases (Generating sentences using keywords: 140, Recommending next sentences: 294) to understand when participants chose to use the sentences generated by each LLM feature. To this end, we extracted the relative positions when the participant chose LLM-generated sentences. (*i.e.*, 0 is the starting position, and 1 is the end position of the journal) Results from the t-test showed that our participants more frequently used sentences from Generating sentences using keywords (M = .27, SD = .31) in the earlier phase of writing an entry than using sentences from Recommending next sentences (M = .53, SD = .27) ($t(432) = 9.59$, $p < .001$) (See Figure 3). Similarly, in the qualitative analysis, many participants responded that they used Generating sentences using keywords at the beginning of the journal to initiate writing. *"I think the most challenging part is starting the first sentence of the diary. I liked that it was easy to generate starting sentences using just a few keywords."* (P24, 21(F)). Another difference in the usage between the two features was that while participants tended to use Generating sentences using keywords when they had a clear idea of what they wanted to write, they turned to Recommending next sentences when seeking a broader perspective or unexpected suggestions.

We then wanted to know what kinds of LLM-generated sentences people brought into their diaries. DiaryMate suggests up to 5 sentences upon the user's request, and the user can select a sentence among them. We analyzed the characteristics of the sentences selected by the participants in the study. We classified the sentences selected by users into five themes as follows: Description of situations, Expression of emotional responses, Reflection, and Lesson learned, Mindsets and Future events and hopes (See Table 2).

5.1.2 Did the use of the LLM feature affect the journal length? We examined the impact of LLM features on the length of participants' journals. The length of a journal can serve as a data point that can help understand the extent of participants' self-disclosure and expression. First, a correlation analysis was conducted to determine how the number of LLM feature requests (*i.e.*, Generating sentences using keywords, Recommending next sentences) affected the participants' journal length. In the case of Recommending next sentences, there was a moderately positive correlation with the length of the journal (r = .5797, p = .003), but in the case of Generating sentences using keywords, there was no significant correlation (r = −.2479, p = .242). We wondered if the journal was getting long simply because the participants used LLM-generated sentences. Therefore, we further investigated whether there was a correlation between the number of sentences participants selected (brought) from the LLM's suggestions and journal length. Interestingly, there was no significant correlation between journal length and the number of sentences chosen to use from the LLM suggestions (Generating sentences using keywords: r = -.2253, p = .2898, Recommending next sentences: r = .1886, p = .3775) (See Figure 4). In summary, these results imply that the more users browse the LLM-generated sentences from Recommending the next sentences, the more likely they will write. However, the number of sentences participants incorporated into their journals was not correlated with the journal length. In other words, the increased length of the journal wasn't merely due to the convenience of having text written on their behalf.

5.1.3 Does the existing habit of keeping a journal affect the use of LLM. Next, we wondered whether the LLM feature was used frequently, even for those who regularly kept a journal. To this end, we examined the difference in the use of LLM features between those with a journal-keeping habit and those without. Through correlation analysis, we found no significant correlation between the participants' existing journal-keeping habits and the use of LLM features (Generating sentences using keywords: r = -.1398, p =

[3] Korean uses a unique, combinatory script, resulting in a lower character count compared to English. Additionally, the agglutinative nature of Korea, particularly its use of particles, makes word counting challenging.

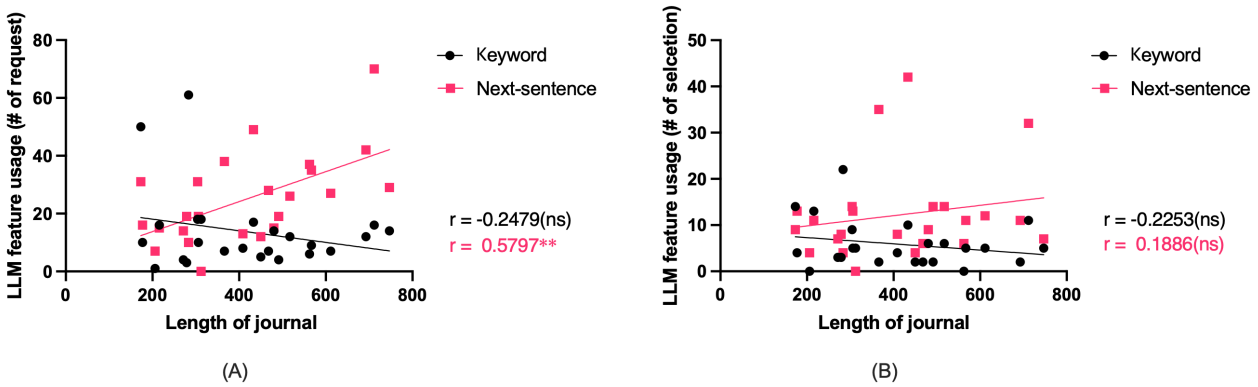

Figure 4: Scatter plots representing the correlation between the length of the journal and the usage of two LLM features: Generating sentences using keywords, and Recommending next sentences. 'r' = Pearson's correlation coefficient. Statistically significant results are reported as *: $p < .05$, **: $p < .01$, ***: $p < .001$

.5148, Recommending next sentences: r = -.2223, p = .2965). In other words, the participants used LLM regardless of their journal-writing proficiency. Through qualitative analysis, we could understand how LLM was useful to those who used to keep their diaries and those who were new to journaling. First, those who write in their diary regularly evaluate the fact that they can reflect on their feelings through LLM sentences. *"People who regularly write in their diary want to understand more about their emotions and moods. In that sense, looking at the sentences recommended by DiaryMate and comparing them was apparently a better way to reflect on my internal feelings."* (P05, 21(F)). On the other hand, those who did not write a diary highly evaluated the LLM function as a tool that provides practical help related to writing, such as starting the first sentence or expressing abstract emotions in writing as mentioned in [44].

Moving on to Section 5.2, we delve deeper into our qualitative analysis, exploring the ways in which participants employed these LLM-generated sentences in their journaling. Additionally, we discuss the perceived advantages and challenges they encountered when utilizing LLM during the journaling process.

5.2 How Do Participants Perceive and Use LLM when Journal Writing?

In this section, we report the qualitative findings on how our participants perceived and used LLM in their journaling practice. In particular, we report both the positive and negative feedback that our participants noted in this field study. To this end, we analyzed in-situ experience and interview data. Regarding the positive aspect, our participants used LLM to re-examine their past feelings and thoughts with diverse perspectives and obtain empathy and comfort from LLM-generated sentences. However, the use of LLM occasionally interrupted the natural progression of a writer's thoughts and inadvertently led them to align more closely with the emotions and sentiments expressed in the LLM-generated text. Users tended to be more influenced by the emotions and atmosphere provided by the LLM, leaning their expressions in that direction, rather than their personal evaluations or thoughts.

5.2.1 Using LLM to revisit past feelings and thoughts from diverse perspectives. Overall, our participants perceived that LLM provides diverse suggestions of what other perspectives could exist about their situation and thoughts: *"Reading the sentences recommended by DiaryMate seemed like it showed the lives of 5 different people and their ways of thinking."* (P23, 21(F)) Although the language model did not show the response from a real person (*e.g.*, crowd, peer), participants recognized it as a valid opinion or a thought that real people would have. The main reason people took the output of LLM meaningfully was that they had a perception that the contents of LLM were generated based on learning from numerous articles of real people: *"I knew that DiaryMate was AI, but in fact, the basis of this AI was written by a human, so I accepted the generated sentence as if a human wrote it."* (P04, 21(M)) Based on this perception, one participant even used the LLM feature to test whether their thought was common or not: *"I was so curious about whether my thoughts were common, so I put them in a keyword and then tested them to see if they came up with the sentence I wanted."* (P15, 22(F))

How did this perception of LLM influence the participants' journaling and retrospection? Oftentimes, it is challenging for a person to objectively examine their feelings and thoughts in a problematic situation. In this regard, our participants compared their thoughts and emotions with the perspectives of other people (suggested by LLM) to re-visit their past emotional/mental states: *"I write a journal because I want to see myself more objectively, so I think it was a meaningful experience to see what other people think about the exact keywords in the same or similar contexts."* (P10, 31(F)) This re-visit also helped them to re-frame past problematic situations in a positive way and make resolutions on how to behave in a similar situation. Furthermore, participants explored their past emotions and thoughts that they could not articulate or discriminate by browsing the lists of sentences and expressions generated by the LLM. For example, P16 mentioned that the LLM-suggested sentence makes subconscious thoughts come to mind: *"I regularly feed street cats as a volunteer. While writing it in my journal, I received a sentence from AI: 'I hope cats live a long and healthy life.' At that moment, I realized that I had the same thought. This was a memorable moment. I think*

Table 2: The types of sentences that participants brought into their journal

Theme	Writings on	Example sentences of LLM sentences
Description of situations	Facts, activities, and events	*"I tried to cut back on drinking because I had to get up early tomorrow and go to Seoul, but I ended up binge drinking."* *"In the end, I had no choice but to take another shower. And I put all the wet laundry in the washing machine and ran it"*
Expression of emotional responses	Thoughts, feelings, emotions about an experience	*"It's raining all day today. The rain is falling and so am I."* *"It was the happiest birthday in my life! I got so many messages and wishes. I was grateful to be with my beloved family."*
Reflections and lesson learned	What writer realized, reflected and learned from experiences	*"I feel so stupid. I blamed myself and regretted being so emotional."* *"It just occurred to me. If life were a clock, what time is it for me when everything has been dull since I was 20?"*
Mindsets	Desires, resolutions on changing one's mindset	*"I promised myself to always stick to the basics on whatever I do."* *"I feel so gloomy all day, even now as I'm writing. But since tomorrow I will start a new week I should cheer up."*
Future events and hopes	Behavioral planning (*e.g.* schedules/events) or hopes for the future	*"On this Sunday I'm planning to go out with my friends."* *"I'm planning to wake up early in the morning tomorrow, wash, eat breakfast, and leave right away."*

there was a part that made me realize a little bit of the thoughts that I could not realize." (P16, 23(F))

5.2.2 Obtaining empathy and comfort while writing using the LLM. Even though our system does not position LLM as a conversational agent (*e.g.*, chatbot), participants perceived LLM as someone who read and responded to their writing: *"In DiaryMate, rather than feeling like I am organizing my thoughts on my own, it seems like writing while having a real conversation with someone."* (P12, 25(F)) Our participants expressed their ambivalent desire to keep the journal as a personal and private record, but simultaneously, they wanted someone who read and responded to their writing. In this view, the participants rated their interaction with LLM positively. For example, P16 felt that LLM's output provided a feeling of comfort and empathy: *"When I used the Recommending next sentence function, the AI usually suggested to me some related feelings about the events I had experienced. I was comforted by the AI because I felt like I was not the only one who had been through it."* (P16, 23(F)) Furthermore, many participants liked the fact that it was a machine and not a person who read their journal. This characteristic (of being read by a machine rather than a real person) encouraged users to write more honestly in a more comfortable manner without worrying about evaluations of others: *"In DiaryMate, the person who listens to me is not human but artificial intelligence, so I was able to write a little more honestly."* (P04, 21(M))

5.2.3 Interactions with LLMs that disrupt the flow of thought. Participants also mentioned some adverse effects of the interaction with LLM on the journal writing process. One of the most representative cases is that using LLM functions could cut off the flow of thought in the writing process. In particular, this interruption of flow mainly occurred while using the Generating sentences using keywords feature and when the LLM showed the sentences sequentially. For example, P01 mentioned that after requesting an AI feature, he felt as if the flow of his thoughts had stopped while simply staring at the text being generated: *"This is what I thought was negative (...) Earlier, when I used to write a journal, I used to constantly organize my thoughts and think about what I wanted to write. However, when I clicked on AI features and sentences started to pop up, I found myself not thinking about anything."* (P01, 23(F)) Such disruption also occurred when an additional task for the LLM feature was required, such as thinking about keywords in generating sentences using keywords: *"Now that I have to come up with keywords, I sometimes felt that the flow of writing a journal was cut off when coming up with appropriate keywords."* (P04, 21(M)) Furthermore, LLMs sometimes interfered with the further inner exploration of human writers by trying to finish the journal: *"AI sometimes tries to stop my thinking process For example, I wanted to go deeper, but AI suggested the sentences as if I was about to end my writing that day."* (P15, 22(F))

5.2.4 Biased towards feelings and emotions suggested by the LLM. Some participants noted that the overall mood and flow of emotions seemed to have been unwittingly influenced by the LLM's output. Because it is difficult to determine an individual's mental/psychological state, people often tend to trust the output of the algorithm rather than their judgment [23]. Similarly, while writing a journal, at some point, participants looked back at their writings and asked, 'Did I really feel that way?': *"Before I started writing my journal, I had thoughts and emotions. But after I read some suggestions from AI, I forgot my feelings and followed what AI said."* (P15, 22(F)) In particular, participants reported similar evaluations when

they looked back at the completed journal. One participant who compared journals stored in DiaryMate and regular journals mentioned that looking at DiaryMate's journal, he sometimes felt as if *"It seems like it's someone else's story."* (P10, 31(F))

5.3 How Do Participants Want LLM to Behave in Journal Writing?

5.3.1 How do participants want to control the LLM behavior? Based on the qualitative data from the interviews, we were able to construct findings on how participants wanted to control the LLM behavior. After using an LLM for ten days, the participants understood to some extent the types of sentences they could obtain from the LLM. Based on this understanding, they shared their wishes on how they wanted to control the language model's behavior suitable for their journal writing situation.

Objectivity and subjectivity. First, the participants noted that they wanted to control whether the LLM would generate sentences about objective descriptions of facts and situations or subjective descriptions of feelings and reactions: *"When I write a journal, I mainly write about events that happened a bit too much, and sometimes I want to add feelings. So, I think it would be nice if I could choose between objective event detail and subjective feeling from the AI"* (P20, 26(F)) Similar to P20, other participants wanted to write a journal wherein the objective facts and subjective feelings were balanced. To this end, they wanted to adjust the nature of the generated sentences according to their needs and intentions. Moreover, similarly, some participants noted that they wanted to adjust the characteristics of LLM-generated sentences between rational and emotional: *"I wish that there was an emotional-rational axis as a parameter that I could control. Sometimes, I want to hear a rational solution for my current situation. And if I need emotional comfort, I want to get emotional consolation and empathy from AI."* (P16, 23(F))

Emotions and mood implied in the LLM-generated sentences. Second, the participants noted that they wanted to control the overall emotions of the sentences generated by the LLM. Some participants mentioned that they wanted to set the overall feeling that they wanted to emphasize in the journal. From that perspective, they mentioned wishes that they could choose a certain range of feelings or emotions that the LLM generated: *"I do not know if this can be called a parameter, but I think it would be nice to be able to select emotions. For example, if I choose the feelings of the day, I will recommend sentences while maintaining a general tone of the emotions I choose."* (P10, 31(F)) Another participant wanted to control the LLM's behavior by describing a person's characteristics: *"I think it would be nice to be able to control the personality of AI in the form of a virtual character. For example, let us say I want to obtain advice from a warm-hearted person. I want to set things up like an age in their 20s, a person with a warm personality."* (P04, 21(M))

5.3.2 What role do participants expect LLM to play in journal writing? In the early stages of the study, most of our participants expected the LLM to read their minds and recommend appropriate sentences to them. In other words, participants initially had high expectations toward an LLM. However, as time progressed, they realized that the LLM was not meant to represent their feelings, but rather to add new perspectives and variety to journal writing: *"When I write in my journal alone, it is almost the same every day. So, for example, what I did today was blah blah. However, there is so much data in AI that it just comes up with unimaginable sentences, so it was meaningful that I could get some fresh direction based on those sentences."* (P21, 23(M)) In this regard, some participants described the role of LLM as a book or dictionary that they could refer to when writing a journal: *"I have used it as a dictionary. If I do not know what sentence to write next, I get a recommendation from AI and reference it."* (P05, 21(F))

Our participants also mentioned that the initiative in the journal writing process should be their own. They wanted the human writer's role to the that of the main character who decides the overall context, flow, and topic, and that of the LLM to be an assistive role that helps the expression or further description. To prevent such a situation where human writers lose initiative, they try to maintain their ownership in the journal writing process. We found unique user behaviors that reflected this desire in the field study. For example, P16 used a citation mark (*e.g.*, "text") when using LLM-generated sentences. In the follow-up interview, she mentioned that she wanted to separate what she wrote from AI to retain her journal as her own writing: *"If I take a sentence made by AI as it is, it feels like copying, in which case, I do not feel like I wrote this journal."* (P16, 23(F))

6 DISCUSSION

In this study, we investigated how our participants embraced and perceived LLM's text-generation capability in a journal writing context. To this end, we designed and deployed the DiaryMate system as a technology probe to collect users' AI-mediated writing experiences and envision future design considerations. In this section, first, we present our general reflections on the findings. Subsequently, we discuss design considerations when exploiting LLM technology in a journal writing context.

6.1 Opportunities and challenges of LLM for supporting users' journal writing

In our study, participants exhibited positive engagement during the journal writing process when assisted by the LLM. This aligns with findings from previous research [3, 10, 17, 29] where the participants of the studies valued the diverse structures of the sentences provided by the LLM, evaluating them as both valuable and stimulating. More specifically, in creative writing contexts, the sentences generated by the LLM often acted as springboards for inspiration, introducing new characters, painting vivid scenes, or subtly shifting a narrative's tone [29, 46].

In a similar vein, our study revealed that participants harnessed such LLM suggestions to enhance their journal entries. Specifically, we identified that the diversity offered by the LLM can facilitate writers in exploring their own thoughts and emotions more deeply. Within DiaryMate, participants primarily documented their past experiences, capturing inner feelings and emotions that can sometimes be ambiguous and hard to articulate. During such introspective writing, they incorporated sentences from the LLM, allowing them to perceive their past from the varied perspectives suggested by the model.

Overall, our findings suggest that interacting with the LLM can amplify the benefits of journaling. The LLM nudges participants to delve deeper into self-expression, uncover novel interpretations and significance from past incidents [45], and potentially influence their future mindset [1].

In our study, participants often viewed the LLM as a collaborative writing partner that could read and respond to their entries. This perception highlighted a unique dichotomy: while participants desired the privacy of their journals, they also yearned for feedback and acknowledgment. DiaryMate addresses this need, offering a platform where users can pen their thoughts privately, yet receive feedback from someone else. This concept aligns with prior research indicating that virtual agents can encourage individuals to express themselves in a more open manner [36]. The non-human nature of the LLM offers an added advantage; users might feel more at ease discussing vulnerabilities without the fear of damaging their reputation, as supported by findings on the hesitancy to disclose personal issues to humans due to potential stigmatization [18].

On the other hand, our study also highlighted potential drawbacks of using the LLM in the writing process, particularly concerning user autonomy [3] and their perception of journal entries. We found that users might lean heavily on LLM's suggestions to articulate past feelings and moods, sometimes sidelining their own assessments. This pattern echoes the observations from reflection studies that leveraged AI, where individuals might override their personal evaluations in favor of AI-generated content [23].

This phenomenon underscores the need for careful consideration of applying LLM in a journal writing context. This is especially because the process of reminiscing and reflecting on the past can have a negative impact on one's mental health. While reflection – the deep thinking and consideration of one's thoughts, experiences, and actions – is known to significantly benefit mental health, rumination, which involves a continuous focus on personal problems, shortcomings, and past failures, is negatively associated with psychological well-being [21]. Considering previous research that suggests opinionated LLMs can steer people's perspectives and viewpoints [25], it's important to consider the possibility that LLMs could also influence users' perspectives and approaches in reflecting on their past.

Our study reveals that LLMs can emulate a therapist's role by providing empathy and comfort to users sharing personal difficulties [32]. However, if users perceive LLMs as actual therapists, this raises ethical and psychological concerns. The empathy from LLMs differs significantly from that offered by professional human counselors. Overlooking this distinction risks users receiving misleading advice, as LLMs cannot fully comprehend or convey the complexity of human emotions, potentially leading to misunderstandings or inadequate expression of feelings.

In our system, even without explicit elements of counseling or therapy, users perceived the LLM as a therapist, influenced by its natural understanding and fluent responses to their inputs. This underscores the importance of clearly communicating the limitations and appropriate usage of LLM technology in personal and emotional contexts. It is vital for designers to grasp the user's expectations when interacting emotionally with LLMs and to offer suitable guidance and support, ensuring users understand the nature and scope of the LLM's capabilities.

Overall, our study emphasizes the need for careful consideration in defining the suitable roles and boundaries of LLM use, particularly in personal and reflective contexts.

6.2 Design Considerations for LLM-mediated Journal Writing

6.2.1 Mitigating over-reliance on LLM outputs. Some of our participants had a basic understanding of the data-driven nature of LLMs, particularly those nine who had completed an AI course. With this background, they tended to interpret the sentences generated by LLMs as if they were crafted by humans, attributing to them a level of meaning and intention one might expect from real individuals. We speculate that this perception arose partly because the sentence-generation capabilities of contemporary LLMs have advanced to a level where distinguishing between LLM and human-created content is challenging. However, some participants attributed to the LLM's outputs the sort of meaning or response they believed a typical person might exhibit in response to specific events. This tendency was further emphasized when they used the LLM's outputs as reference points for introspection. We argue that it is crucial for systems to actively educate users about the inherent limitations and potential inaccuracies of language models, ensuring users don't credit undue significance or trust to the generated content from LLM.

Furthermore, contrasting AI systems with predictable responses, the multifaceted and varied outputs of the LLM pose challenges in determining the range of their responses in advance. Additionally, it remains challenging to identify and fix biases within LLM outputs consistently [2, 14]. Compared to fiction and other usual writing styles, when LLMs give suggestions for personal journals, it can really affect how the person feels and thinks about themselves. Consequently, design considerations should prioritize measures that prevent users from placing undue credibility and significance on language model outputs. For instance, we could think about implementing features within the system that periodically inform and remind users about the limitations and potential inaccuracies of LLMs. This could be in the form of tooltips, warning messages, or educational modules that explain how LLMs generate content and why it should not be interpreted as advice or wisdom from a human perspective.

6.2.2 Supporting users' exploration of their inner self. Our study showed that when people wrote in their journals using the LLM, they found it more helpful to explore different thoughts and ideas from the LLM than just looking for the "best" sentence to use. When writing, participants liked comparing many LLM suggestions to their own feelings. They said it helped them see their feelings from a new angle. This process enabled them to view their feelings from diverse perspectives, thereby enriching their self-reflection process. We observed that exposure to a wider range of LLM suggestions, even if not all were used, led to more extended journal entries. This suggests that providing a variety of viewpoints, including polarizing ones, might foster deeper introspection and self-discovery. The idea behind introducing polarizing perspectives in an LLM is to encourage users to engage more deeply with their thoughts and emotions. By comparing and contrasting different viewpoints, users are prompted to reflect more critically and thoroughly on

their inner selves. This can lead to a more nuanced understanding of their emotions and experiences. For instance, if a user is journaling about a specific event, the LLM might offer one suggestion that interprets the event positively and another that offers a more critical or negative viewpoint. The user is then encouraged to reflect on these different perspectives, which can help them explore their feelings and thoughts more fully, leading to a richer, more complex process of self-reflection. Furthermore, this approach not only widens the scope for self-reflection but also prevents a unidimensional interpretation of one's emotions, influenced solely by the LLM's output.

6.2.3 Assuring user agency in journal writing. Consistent with user perceptions of human-AI collaborative tasks shown in a previous study [34], our participants also emphasized that human writers should take the lead in journal writing. Moreover, we could identify the desire for controllability so that the participants could adjust the characteristics and behavior of the language model to their needs and intentions, which revealed that participants wanted to take the lead in the process of journal writing. Although concerns about autonomy and ownership have already been addressed in the human-AI collaborative writing research domain, in this study, we further emphasize that system design for achieving user initiative in journal writing should be prioritized. First, the user's autonomy and sense of agency are critical elements in the process of looking back on past events and emotions and finding meaning within them [7]. Particularly, it has been reported that writing with LLMs possessing opinions can significantly steer people's perspectives and viewpoints [25]. In our study, we additionally identified the possibility that users could easily follow the algorithm's output rather than their judgment, especially when expressing emotions and feelings that are difficult to discern and describe objectively [23]. In a similar way, some participants had written journals that deviated from their initial intentions and plans. Therefore, we suggest that when providing the LLM output containing personal emotions and feelings to the user, we might consider designing an interface that allows users to review and evaluate LLM suggestions before accepting them. This ensures that users consciously decide whether the suggested content aligns with their thoughts and feelings. Alternatively, we could consider a design where the role of the LLM is not to write the content of the diary directly but rather to help the user think about what to input and what topics to consider. For example, the LLM could generate reflection questions, thereby aiding the user's self-led retrospection [28]. This can help users maintain a critical stance and ensure that their journal entries align with their original intentions and thoughts.

6.3 Limitations and Future work

Our study has several limitations. There is a possibility of sampling bias and issues of generalizability because we recruited university students as participants. For example, university students may have different levels of AI literacy than others. Future studies could compensate for this issue by recruiting participants, considering their background knowledge of AI. Next, we acknowledge that 10 days may not be sufficient for examining long-term engagement. Nevertheless, compared with several LLM co-authoring studies that were mainly conducted in a single session, we used an approach that allows people to interact with LLM in the wild. We will continue to explore the issue of human-AI interactions with an improved system to understand how people's acceptance and perception of LLM change in the longer term.

7 CONCLUSION

In this study, we addressed the opportunities and challenges of utilizing LLM technology to support users' journal writing by designing DiaryMate, a technological probe to allow users to interact with LLM in daily journal writing. Through a 10-day field deployment study using DiaryMate, we collected qualitative and quantitative data on how people perceive and adapt LLM technology to their personal journaling processes. Participants positively rated the features and experiences provided by LLM technology, and the use of LLM technology helped enrich their journaling experience. Participants used the diversity of the sentences generated by the LLM as a tool to revisit their past experiences from various perspectives. They also perceived LLM as an emotional partner who listened and responded to their personal stories. However, at the same time, we found that people gave excessive meaning and credibility to sentences generated by the LLM, using them as a reference point for reflection and often prioritizing the expression of emotions by the LLM over their own. Based on the findings, we provide a design consideration that prevents users from overvaluing LLM's sentences, enhances the experience of exploring diverse perspectives, and provides the initiative to users in self-awareness. We hope our research will provide insights and research agendas for designing interactions in AI-mediated writing.

ACKNOWLEDGMENTS

We thank our study participants for their time and efforts. This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT; No. 2022-0-00064, Development of Human Digital Twin Technologies for Prediction and Management of Emotion Workers' Mental Health Risks) and National Research Foundation of Korea (NRF; RS-2023-00262527). Additionally, we extend our gratitude to NAVER CLOUD and the National IT Industry Promotion Agency (NIPA) for providing us with the HyperCLOVA API which made this research possible.

REFERENCES

[1] Nancy Allison. 1999. *The illustrated encyclopedia of body-mind disciplines.* Taylor & Francis.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. *Designing Interactive Systems Conference* (2022), 1209–1227. https://doi.org/10.1145/3532106.3533506

[4] David Boud. 2001. Using journal writing to enhance reflective practice. *New directions for adult and continuing education* 2001, 90 (2001), 9–18.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] Hronn Brynjarsdottir, Maria Håkansson, James Pierce, Eric Baumer, Carl DiSalvo, and Phoebe Sengers. 2012. Sustainably Unpersuaded: How Persuasion Narrows Our Vision of Sustainability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 947–956. https://doi.org/10.1145/2207676.2208539

[8] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B. Chilton. [n. d.]. How Novelists Use Generative Language Models: An Exploratory User Study *(HAI-GEN+user2agent@IUI 2020)*. An early exploratory study found that autocompletefrom a language model did not provide enough control for novelist.

[9] Ke-Li Chiu and Rohan Alexander. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407* (2021).

[10] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[11] Kenneth W. Church and Lisa F. Rau. 1995. Commercial Applications of Natural Language Processing. *Commun. ACM* 38, 11 (nov 1995), 71–79. https://doi.org/10.1145/219717.219778

[12] Robert Dale. 2016. Checking in on grammar checking. *Natural Language Engineering* 22, 3 (2016), 491–495.

[13] Robert Dale and Jette Viethen. 2021. The automated writing assistance landscape in 2021. *Natural Language Engineering* 27, 4 (2021), 511–518. https://doi.org/10.1017/s1351324921000164

[14] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 862–872.

[15] Raymond Fok and Daniel S Weld. 2023. What Can't Large Language Models Do? The Future of AI-Assisted Academic Writing. In *In2Writing Workshop at CHI*.

[16] Katy Ilonka Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. 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, 1–12. https://doi.org/10.1145/3290605.3300526

[17] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for Science Writing using Language Models. *Designing Interactive Systems Conference* (2022), 1002–1019. https://doi.org/10.1145/3532106.3533533

[18] Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and schuster.

[19] Amy L. Gonzales, Tiffany Y. Ng, OJ Zhao, and Geri Gay. 2010. Motivating Expressive Writing with a Text-to-Sound Application. 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, 1937–1940. https://doi.org/10.1145/1753326.1753618

[20] Seung-Soo Ha and Seok-Man Kwon. 2016. Efficacy of the Strengths-based Writing Intervention among the Clinical Adolescents with Externalizing Maladjustment Behaviors. *Korean Journal of Clinical Psychology* 35, 1 (2016), 139–163. https://doi.org/10.15842/kjcp.2016.35.1.008

[21] Rick Harrington and Donald A Loffredo. 2010. Insight, rumination, and self-reflection as predictors of well-being. *The Journal of psychology* 145, 1 (2010), 39–57.

[22] Elisa L. Hill-Yardin, Mark R. Hutchinson, Robin Laycock, and Sarah J. Spencer. 2023. A Chat(GPT) about the future of scientific publishing. *Brain, Behavior, and Immunity* 110 (2023), 152–154. https://doi.org/10.1016/j.bbi.2023.02.022

[23] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 114 (sep 2018), 31 pages. https://doi.org/10.1145/3264924

[24] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology Probes: Inspiring Design for and with Families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/642611.642616

[25] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3544548.3581196

[26] Matt Jones, Philippe Palanque, Albrecht Schmidt, Tovi Grossman, Kevin Huang, Patrick J Sparto, Sara Kiesler, Asim Smailagic, Jennifer Mankoff, and Dan Siewiorek. 2014. A technology probe of wearable in-home computer-assisted physical therapy. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 2541–2550. https://doi.org/10.1145/2556288.2557416

[27] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, et al. 2021. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. *arXiv preprint arXiv:2109.04650* (2021).

[28] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 2, Article 70 (jul 2018), 26 pages. https://doi.org/10.1145/3214273

[29] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. *CHI Conference on Human Factors in Computing Systems* (2022), 1–19. https://doi.org/10.1145/3491102.3502030 arXiv:2201.06796

[30] Haiwei Ma, C. Estelle Smith, Lu He, Saumik Narayanan, Robert A. Giaquinto, Roni Evans, Linda Hanson, and Svetlana Yarosh. 2017. Write for Life: Persisting in Online Health Communities through Expressive Writing and Social Support. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 73 (dec 2017), 24 pages. https://doi.org/10.1145/3134708

[31] John D Mayer, David R Caruso, and Peter Salovey. 1999. Emotional intelligence meets traditional standards for an intelligence. *Intelligence* 27, 4 (1999), 267–298.

[32] Sharon Myers. 2000. Empathic listening: Reports on the experience of being heard. *Journal of Humanistic Psychology* 40, 2 (2000), 148–173.

[33] Changhoon Oh, Jinhan Choi, Sungwoo Lee, SoHyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding User Perception of Automated News Generation System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376811

[34] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174223

[35] R OpenAI. 2023. GPT-4 technical report. *arXiv* (2023), 2303–08774.

[36] SoHyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. 2021. "I wrote as if I were telling a story to someone I knew.": Designing Chatbot Interactions for Expressive Writing in Mental Health. *Designing Interactive Systems Conference 2021* (2021), 926–941. https://doi.org/10.1145/3461778.3462143

[37] James W Pennebaker. 1985. Traumatic experience and psychosomatic disease: Exploring the roles of behavioural inhibition, obsession, and confiding. *Canadian Psychology/Psychologie canadienne* 26, 2 (1985), 82.

[38] Tara Riddell, Jane Nassif, Ana Hategan, and Joanna Jarecki. 2020. Healthy Habits: Positive Psychology, Journaling, Meditation, and Nature Therapy. *Humanism and Resilience in Residency Training: A Guide to Physician Wellness* (2020), 439–472.

[39] Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality* 9, 3 (1990), 185–211.

[40] Michele Salvagno, Fabio Silvio Taccone, Alberto Giovanni Gerli, et al. 2023. Can artificial intelligence help for scientific writing? *Critical care* 27, 1 (2023), 1–5.

[41] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (2023), 46–57. https://doi.org/10.1038/s42256-022-00593-2

[42] Joshua M Smyth. 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. *Journal of consulting and clinical psychology* 66, 1 (1998), 174.

[43] Daniel H Solomon, Kelli D Allen, Patricia Katz, Amr H Sawalha, and Ed Yelin. 2023. ChatGPT, et al... Artificial Intelligence, Authorship, and Medical Publishing. *ACR Open Rheumatology* 5, 6 (2023), 288.

[44] Cheryl Travers. 2011. Unveiling a reflective diary methodology for exploring the lived experiences of stress and coping. *Journal of Vocational Behavior* 79, 1 (2011), 204–216. https://doi.org/10.1016/j.jvb.2010.11.007

[45] Philip M. Ullrich and Susan K. Lutgendorf. 2002. Journaling about stressful events: Effects of cognitive processing and emotional expression. *Annals of Behavioral Medicine* 24, 3 (2002), 244–250. https://doi.org/10.1207/s15324796abm2403_10

[46] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. 2022. AI as an Active Writer: Interaction strategies with generated text in human-AI collaborative fiction writing. In *Joint Proceedings of the ACM IUI Workshops 2022*,

Vol. 10.
[47] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. *27th International Conference on Intelligent User Interfaces* (2022), 841–852. https://doi.org/10.1145/3490099.3511105