TellTime: An AI-Augmented Calendar with a Voice Interface for Collecting Time-Use Data

Michael J Hoefer
Department of Computer and Information Sciences
University of St. Thomas
Saint Paul, Minnesota, USA
michael.hoefer@stthomas.edu

Raegan Rychecky
Department of Computer Science
University of Colorado Boulder
Boulder, Colorado, USA
raegan.rychecky@colorado.edu

Max Gong
Department of Information Science
University of Colorado Boulder
Boulder, Colorado, USA
max.gong@colorado.edu

Stephen Voida
Department of Information Science
University of Colorado Boulder
Boulder, Colorado, USA
svoida@colorado.edu

Abstract

Reports of daily time use are gathered for both large-scale population-level surveys and individual self-tracking in personal informatics systems. However, gathering a complete self-report record of an individual's daily activities is time consuming and cognitively demanding. Gathering structured time records via spoken narrative can reduce the burden of self-tracking and support natural "storytelling" as a method of data collection. We designed, built, and evaluated TellTime, a voice interface to a calendar using a large language model. Over three days, 18 participants completed the Day Reconstruction Method with three different calendar interfaces: manual-only, voice-only, and hybrid. We analyzed the user experience with surveys, semi-structured interviews, and telemetry data. Our findings show that users prefer hybrid interaction, where they first gather data via a spoken narrative, and then manually make precise adjustments to specific activities. This work suggests that LLMs can enable voice interaction with calendar systems, and that spoken narrative can improve the user experience of gathering self-report time records.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**; **Natural language interfaces**; **User studies**.

Keywords

Human-AI interaction, interactive system design, ai-augmented systems, time-use, large language models, voice interface, self-tracking, quantified self, time studies, calendar, time-use diary

ACM Reference Format:
Michael J Hoefer, Raegan Rychecky, Max Gong, and Stephen Voida. 2025. TellTime: An AI-Augmented Calendar with a Voice Interface for Collecting Time-Use Data. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3708359.3712116

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712116

1 Introduction

Time is a valuable and finite resource, and underlies the basis of our experience. Individuals, governments, and researchers are interested in collecting data on time use. Individuals track their time use to improve productivity [40], track health-related time-use such as sleep duration [8, 38], and to engage in lifelogging more generally [27, 43]. Researchers seek to understand time allocation at various social scales ranging from small participant samples in psychological studies, up to the global population [20], to measure the amount of time spent extracting resources, or dealing with waste, for example [18]. State governments [55] and multinational organizations [17] are motivated to learn how their citizens spend their time to inform policy decisions related to labor and health, for example [28].

The goal of data collection is to construct a complete record of how an individual spends their time, often for a given window of time such as a specific 24-hour day [65]. Self-reports of time use are captured manually using paper diaries, phone interviews, or electronic systems with manual entry. These methods require significant effort from participants (and surveyors). As an example, the American Time Use Survey (ATUS) conducts phone interviews with a sample of households in the US and records a complete record of individual participants' days. The ATUS previously consisted of time records of 3,375 households per month but in 2003 it reduced that number by 35% to 2,190 households per month for cost reasons [55]. For individuals, the effort of self-tracking is a major barrier to collecting time-use data [29, 31], and simply creating electronic versions of the data collection instrument does not reduce this burden [52].

To reduce the effort needed to collect self-report time-use data, we present *TellTime*—a calendar system with a voice interface that supports hybrid human-AI interaction for collecting time-use data. TellTime enables users to simply speak a natural language narrative about their day to the system, which utilizes a large language model (LLM) to parse the narrative into discrete labeled activities.

Users can make modifications to the time record via additional spoken commands or via manual modification (akin to commercially available calendar systems).

Using a research-through-design approach [76], we designed and built three versions of the system: (i) a completely manual interface, (ii) a voice-only AI interface, and (iii), a hybrid manual–AI interface. To evaluate the system, we conducted a user study with 18 participants. In a three-way within-subjects randomized controlled trial (RCT) over three days, participants were tasked to reconstruct a full record of their time use by creating activities on the calendar, using each interface (randomly selected and balanced across participants). We collected system telemetry data, qualitative and quantitative feedback in daily questionnaires, and conducted a closing interview with each participant to understand their experience using the system. We also collected a dataset of 120 transcripts of spoken input to the system, and annotated each instance with the participant's intention(s), the 'gold standard' output, scored each instance with a system 'accuracy,' and conducted an error analysis.

The majority of errors made by the system were placing activities at the wrong time on the calendar, an issue which was exacerbated when participants provided only a single "time anchor" in their spoken narrative. Despite the errors, 16 of the 18 participants preferred to use the AI-voice interface over manually entering events, and most participants utilized a hybrid approach to reconstructing their day—speaking a narrative to the system to create an initial set of activities, and then making specific manual modifications to complete the record. Although the voice interface did not result in significantly faster data collection times, participants perceived that it did. There were no significant differences found across interfaces for average number of activities reported, a measure of quality of self-reported time-use data [35].

Overall, this work makes the following contributions:

- A hybrid approach to human–AI interaction on a calendar
- TellTime, an LLM-augmented voice interface for gathering time-use data
- Improved understanding of situations where LLMs do not perform well at parsing spoken narratives into temporal events
- Design recommendations for using AI to collect time-use data

2 Background and Related Work

Our work contributes to, and is situated in, methods for collecting time-use data, designing personal informatics systems, and AI for self-tracking and time management. We briefly review related literature from these areas.

2.1 Methods for Collecting Data on Time Use

There are a variety of methods for collecting time-use data, ranging from stylized questions [21], direct observation [58], using a time diary [63], *in situ* reporting (such as the experience sampling method) [15], and a hybrid method involving retrospective episodic recall called the Day Reconstruction Method [36].

The experience sampling method (ESM) is often preferred over retrospective methods because the *in situ* reporting reduces recall biases. However, the ESM does not provide a full report of an individual's time use [67], and involves specific hardware that interrupts a participant during their day. Therefore, many time-use studies rely on time diaries [63]. Participants are asked to provide a report of everything they were doing in a 24-hour window, either for "yesterday" (known as a "retrospective" diary) or for "tomorrow" (to be filled out by the participant throughout their day) [67]. Participants either fill out a paper diary, an electronic diary, or, as is the case with the ATUS, participants report their time use via an interview with researchers over the phone.

The method used by the ATUS is similar to the Day Reconstruction Method (DRM), which instructs participants to divide their day up into episodes that represent distinct activities, and report on each episode. The DRM does not require participants to manage a diary, has been shown to have similar levels of accuracy to the ESM [36], and has the benefit of providing a full record of an individual's day. The DRM can be completed in one sitting, making it a cost-effective method for gathering time-use data.

The primary issues with the DRM are recall biases and the burden of completing the task. Technology Supported Reconstruction is an approach to improve recall in the DRM by providing individuals with cues based on passively-logged data gathered throughout the day [37]. For example, the Footprint Tracker provided individuals with visual, geographic, and social cues during recall to trigger specific memories [25]. EmoSnaps captured photos of a participant's face throughout the day, and presented the images to them during recall to improve memory of experienced emotions [54]. While these efforts improve recall, they rely on gathering supplemental data, which may not be available to researchers in traditional population-level time studies.

There remains an open question of how to utilize technology to support the process of participants completing the DRM without using auxiliary data. Electronic calendar interfaces are closely connected to the goal of gathering time-use data, as they present a familiar metaphor and enable thinking of "time-as-artifact" [56]. In real-life usage, calendars are used in a variety of idiosyncratic personal ecologies [16]. The design of existing calendar systems is not specifically focused on supporting the Day Reconstruction Method, but rather "calendar work" generally [56]. Simply creating electronic versions of the data collection instrument is not enough [52], the task of tracking, and the associated burden, remains the same.

Recent work indicates that narrative coherence may facilitate improved recall [10], suggesting that collecting data via a spoken narrative may help in gathering time-use data. Prior work in voice-interactions with a calendar in HCI has been limited to selecting a date from a calendar picker using audio commands [7], reading the contents of a calendar, or creating individual events (not parsing narratives) [1]. The task of extracting temporal information about events from natural language has been studied in the field of natural language processing (NLP), in particular with the TempEval task [68, 70]. Recent advances in large language models (LLMs) motivate designing new methods of gathering time-use data involving spoken narrative, for both population-level studies and personal informatics systems.

2.1.1 *Temporal reasoning in LLMs.*

Several recent studies have sought to benchmark and evaluate an LLM's capability to reason about time and events [9, 64, 71, 74, 75]. Jain et al. [34] evaluate eight LLMs on varying temporal commonsense reasoning tasks, finding that while most LLMs they evaluated performed strongly on tasks related to reasoning about event duration, all but GPT-3.5 and FLAN-T5 performed poorly on event ordering tasks. Notably, for the current work, prior work has found that presenting LLMs with events in order (rather than a random order) helps improve temporal processing [19], making a temporally linear narrative a good use case for LLMs.

2.2 Personal Informatics for Time-Use

Personal informatics systems help individuals gather and reflect on data about their lives, in order to help individuals better "know thyself" [45]. Personal informatics systems have been designed to help individuals in particular domains of life, such as sleep [38], fitness [46], nutrition [13], stress [60], and social activity [51]. Other systems, such as OmniTrack, have been designed for flexibility, enabling users to design custom trackers [41].

In addition to commercial applications like Toggl[1], various personal informatics systems have been designed by researchers in order to help individuals gather data about their time use. TimeAware is focused on helping individuals track their use of time on a computer, and had two interfaces, one that focused on visualizing distractions, and the other that focused on visualizing productive time [40]. MeTime helps people track their computer use via automated or manual tracking method [73].

The ExtraSensory app facilitates collection of time-use data using features of both the ESM and DRM [69]. Activity River is a proof-of-concept system for planning, logging, and visualizing time use [3]; in a role-playing situation, participants found its visualizations helpful for reflection. A technology probe for gathering data on need satisfaction via the DRM supported participants in learning about themselves, but self-tracking was a burden, and participants expressed interest in collecting data via voice [31].

The MyMove system utilized a smart watch and voice interface for collecting activity data (activity, duration, and perceived effort) from elderly persons *in situ* with an ESM method [39]. While MyMove utilized a voice interface to track activity, the voice recordings were not automatically parsed into a structured format for participants' immediate use, but rather saved in a database for later analysis. The study found relatively low word error rates in transcription, and that the participants found it easy to use the voice interface to record their data [39]. MyMove highlights the importance of using a voice interface for differently-abled individuals.

2.3 AI for Self-Tracking and Time Management

Traditional personal informatics (and time tracking) systems tend to place high cognitive burdens on the users, motivating the exploration of using AI to assist in tracking [50]. Kim et al. proposed and explored a new NLP task for structured activity logging from spoken language [42]. Other work utilized an LLM-based chatbot to gather self-reported information about sleep, fitness, nutrition, and work [72]. These studies show that LLMs can be used to extract personal data from natural language self-report.

Other studies seek to use machine learning and AI to provide more automated data capture. For example, one study gathered sensor data from older adults in combination with text and voice annotations to understand stress [22]. Another explored annotations both *in situ* and via recall, with and without visualization of sensor data [32]. Others have sought to reduce the burden of the ESM by predicting activity annotation directly from sensors with machine learning [47]. Generally, these human activity recognition (HAR) efforts focus on automating the data collection process by prediction, rather than using AI to support traditional self-report. These methods tend to require individuals to wear sensors and/or share their personal device data to be analyzed, raising privacy concerns. Another specific risk of automating data collection (in a context of personal informatics) is that participants will simply not engage with the process at all, which was noticed in a study of RescueTime [11].

There have been multiple AI personal assistants developed to help individuals in scheduling meetings via email, such as PTIME [4], CMRadar [53], and Calendar.help [14]. Calendar.help effectively utilized hybrid machine and human intelligence to assist individuals in scheduling meetings via email [14]. This related work shows the importance of using a voice interface to collect data, the effectiveness of hybrid human–AI intelligence for calendar management tasks, and a user preference for using the Day Reconstruction System (or general, non-situated recall [32]) for personal time tracking.

3 TellTime: An AI-Augmented Calendar Voice Interface

The TellTime system is an AI-augmented calendar where the data is elicited completely by the participant, but parsed and structured by the AI to reduce the cognitive burden of self-tracking. The design seeks to support episodic recall of time use by facilitating data capture via a spoken narrative which is then parsed into structured activities on the calendar. The design of the system was guided by the following three design goals which are intended to improve the experience of self-reporting time use without relying on external data:

- Support the traditional Day Reconstruction Method
- Facilitate natural language data collection in both narrative and non-narrative format
- Allow for both manual and voice input, enabling hybrid human–AI interaction

3.1 Hybrid Interaction Approach

TellTime implements a hybrid model of human–AI interaction, allowing both manual input from the user and voice input parsed by the AI. The methods of interaction of the hybrid interface are shown in Fig 1. Both the human and the AI have access to the calendar state and can make modifications to the events on the calendar. The human has the choice whether to manually create and modify events on the calendar or provide natural language input to the AI, which will then make modifications on the human's behalf.

3.2 Interface Description

3.2.1 Manual Interactions. The manual interface seeks to replicate the functionality of existing calendar systems, such as Google Calendar or Microsoft Outlook. Users can click anywhere on the

[1] https://toggl.com/

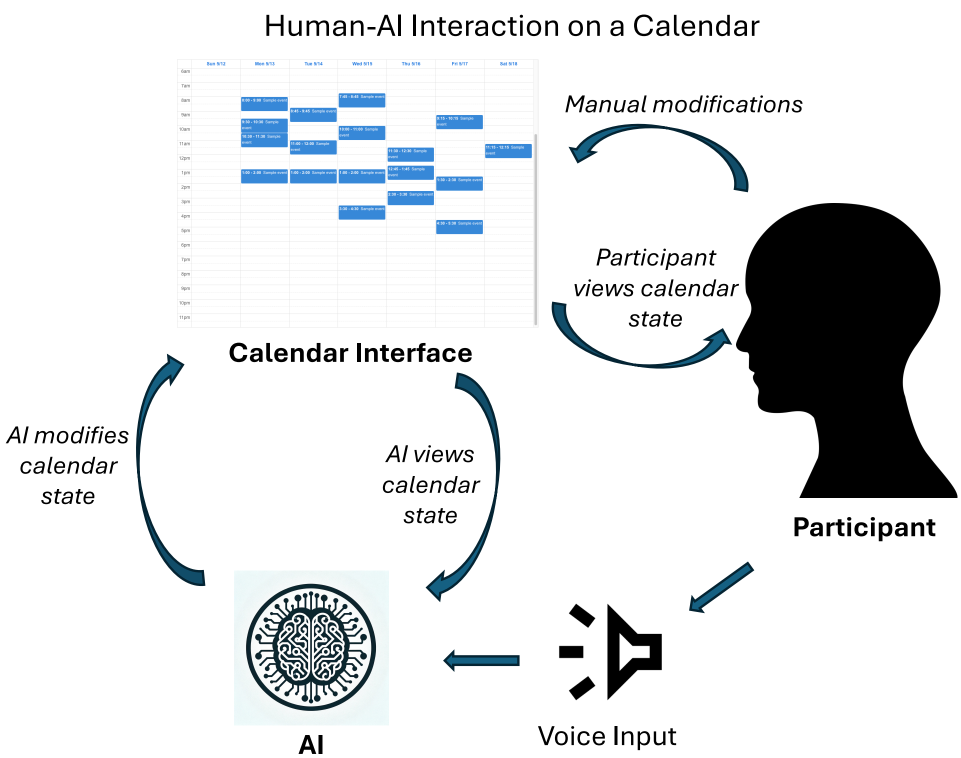

Figure 1: Hybrid interaction includes both manual modification of the calendar state (via mouse and keyboard) as well via voice commands to an AI agent which modifies the calendar state.

calendar to create a new, 15 minute event at that time, or click and drag to specify a duration for the event. An alert prompts the user to enter a title for the event. Users can click and drag on a pre-existing event in order to change the time, and can click and drag on the bottom edge of an event to change the duration of an event. Users can click on an event, which brings up a modal pop-up (Figure 2), where users can re-name the event or manually adjust the time using a time picker.

3.2.2 Voice Interactions. To use the voice interface, users click "Start Recording" to activate the microphone, and then speak a narrative about their day or provide any other natural language command to the calendar, such as requesting additions, modifications, or deletions to events on the calendar. When done speaking, users click "Stop Recording," and the transcript of their speech is combined with a system prompt (see Appendix A for the system prompts) and a comma-separated value string representation of the current events on the calendar (the audio recording is not saved). These data are then sent to the OpenAI API, which returns the modified calendar state. After the event state has been modified by the AI, an "Undo" button becomes enabled, allowing users to undo the most recent action made by the AI. The voice interface and calendar appear in Figure 3.

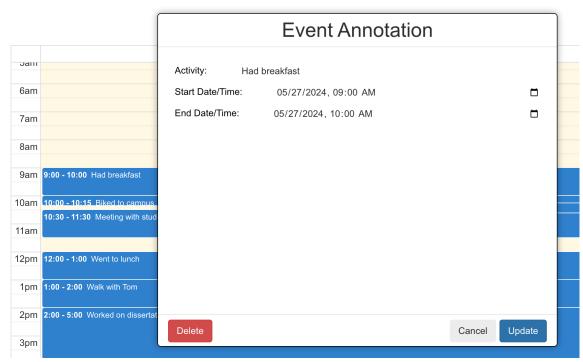

Figure 2: Modal for manual interaction, which appears when an event is clicked on, and allows the participant to change the event title and the start time and stop time.

3.3 System Implementation

TellTime was implemented as a React.js[2] web application utilizing the fullcalendar.io [61] JavaScript package. The FullCalendar package and add-ons provide support for standard calendar interactions, such as calendar event click-and-drag interactions and calendar

[2]https://react.dev/

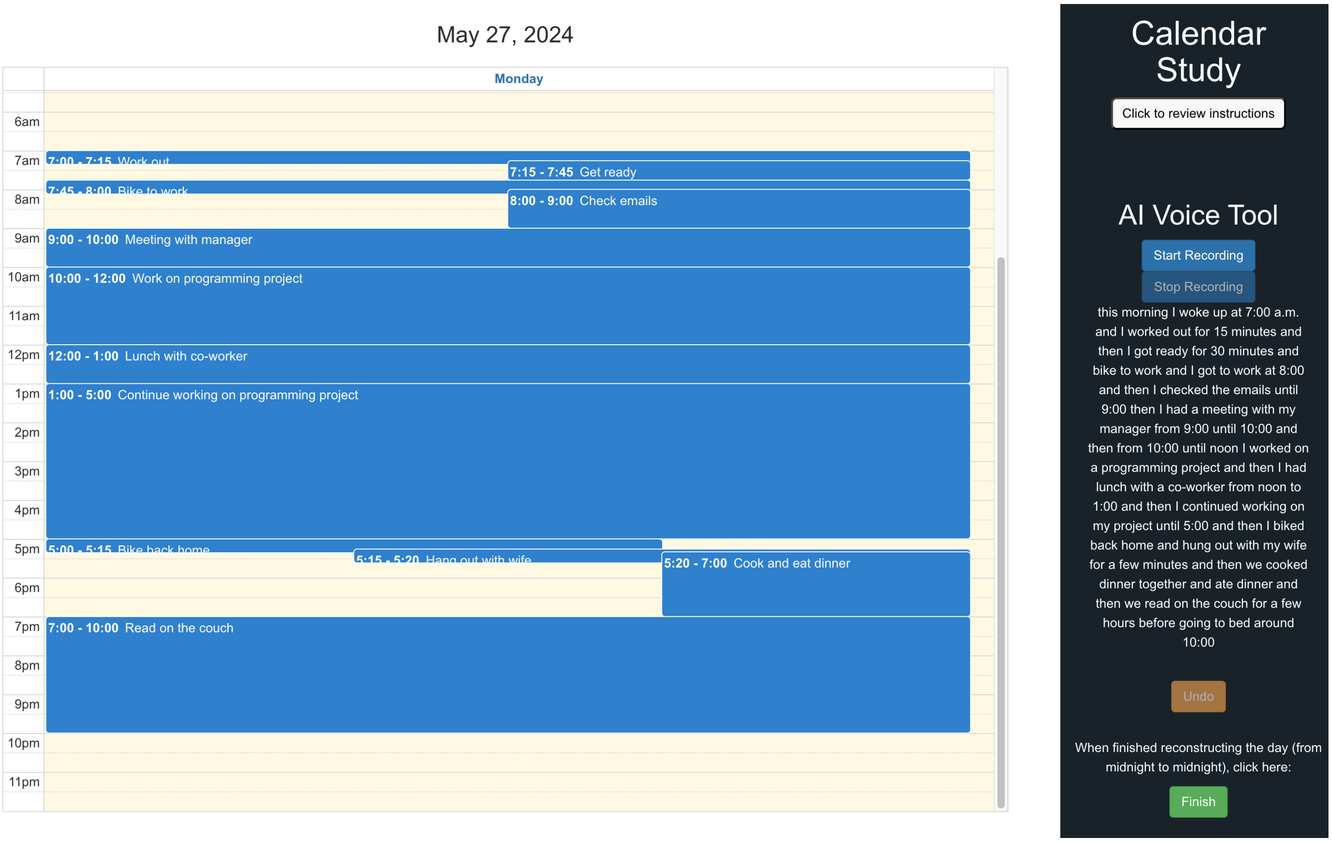

Figure 3: TellTime system with calendar and AI voice interface. Participants can record a narrative of their day, which will be parsed by an LLM into structured events, and shown on the calendar.

state management. We used the WebSpeechAPI[3] to provide a live transcription of the user's spoken narrative, displayed in the sidebar. The transcription, along with the current calendar state and system prompt, is sent to OpenAI's GPT-4[4] large language model via their API.

To develop the system prompts, we created an initial test narrative and experimented with different prompts in the OpenAI Sandbox. We iterated on the prompt until it was able to successfully parse a wide variety of narratives. Initially, we used OpenAI's functions feature to specify a specific JSON format of the LLM response. However, when implementing the functions API call in the calendar interface, the latency was significantly higher than a traditional unstructured API call, making the calendar interface feel less responsive. We were able to achieve similar performance by instructing the LLM to output a series of comma-separated values separated by a newline character.

The system uses two different prompts: one for an empty calendar, the other for when there are pre-existing events on the calendar. The system will parse the output of the LLM into calendar events, which are then displayed on the screen to the user. The final prompts used in the system are in Appendix A.

[3]https://wicg.github.io/speech-api/
[4]Specifically, gpt-4-1106-preview https://openai.com/index/gpt-4/

Telemetry data (timestamps, voice input transcripts, click event details, lists of events, coded identifiers) was sent from the React front end to an AWS DynamoDB database[5] via AWS Lambda function calls[6]. Custom URLs containing participant IDs and interface versions were used to segment participant data. A demo site for the system is available online[7], and the release binaries are available on Github[8].

3.4 Manual, AI-Only, and Hybrid Versions

We developed three different versions of the system: one with only manual controls, one with only a voice interface (AI-only), and one that incorporated both manual and voice interaction modalities. The manual version had no voice control, and all interaction was done with the keyboard and mouse. The AI-only version had disabled manual controls, so the participant could not click to create or modify events on the calendar; all modifications had to be done through the voice interface. The hybrid version allowed the user to make both manual and voice modifications at will.

[5]https://aws.amazon.com/dynamodb/
[6]https://aws.amazon.com/lambda/
[7]A system demo can be found at https://telltime-ai.web.app/. The voice interface requires an OpenAI API key.
[8]Release build is available at https://github.com/mjhoefer/TellTime-AI.

Participant Demographics (n = 18)	
Age	6 (18–24 years) 9 (25–34 years) 2 (45–54 years) 1 (65–74 years)
Gender	8 female (incl. trans f) 8 male (incl. trans m) 2 Prefer not to say
Race	8 Black or African American 7 White 1 Asian 1 Some other race or origin 1 Prefer not to say
Employment	7 student 6 employed in a for-profit organization 4 self-employed 1 employed in a non-profit or government
Income	3 ($10,000–$24,999) 7 ($25,000–$49,999) 3 ($50,000–$74,999) 2 ($75,000–$99,999) 1 ($100,000–$149,999) 2 ($150,000 and greater)

Table 1: Participant demographics (presented in aggregate to preserve identity).

4 Evaluation

In order to explore how different interfaces lead to different participant outcomes, we conducted a comparative user study across the three different interfaces. The study design utilized a within- and between-subject approach where each participant was exposed to each of the three interfaces, albeit in a random order (this order was balanced across participants). Participants reconstructed their day using each interface. The presentation of the interfaces was randomized to control for learning effects due to participants gaining experience reconstructing events from their day.

4.1 Participants

18 participants were recruited via word of mouth, Slack channels, Facebook posts, and via the `r/quantifiedself` subreddit. The criteria for inclusion in the study was that participants be 18 years of age or older, speak English fluently, and have a computer with an internet connection. Participants were provided compensation of $10 for participating in the study, in the form of an Amazon.com gift card. Compensation was prorated based on the activities completed, with $3 earned for the first session, $3 for the second session, and $4 for the final session (which included the closing interview). 14 participants had previous experience with self-tracking, 4 did not. Participant demographics are shown in Table 1.

4.2 Procedure

Participants completed the Day Reconstruction Method on a recorded Zoom meeting with one of three web application interfaces (manual, AI-only, and hybrid) across three separate days (randomly selected and balanced across participants). The overall flow of the study is shown in Figure 4. Participants first completed a background survey which asked them questions about their demographic identity, as well as questions about their previous usage of self-tracking technology. This survey also included the Media and Technology Usage and Attitudes Scale [59] and the AI attitude scale (AI-AS4) [26].

Participants were instructed to think-aloud their thoughts while using the system [44], in order to elicit data on the qualitative user experience. When TellTime is first opened (via a custom link sent to the participant), an initial start-up modal dialog is shown to the participant, with instructions for completing the study (full instructions can be found in Appendix B). When participants click "Start," the dialog box closes, and a hidden timer begins counting up in the background. Participants then reconstruct their day using the controls available to them (manually, with the voice-interface, or both). When participants are finished (re-)constructing their day, they click "Finish," which stops the timer, and opens another modal dialog with a link to a daily questionnaire.

The daily questionnaire included the NASA Task Load Index (NASA-TLX) instrument [30], the System Usability Scale (SUS) instrument [6], free response questions for general feedback, and a 5-point Likert scale for perceived system accuracy. The questionnaire also asked participants, "Without looking at the clock, how much time do you think this exercise took?"

At the end of the third session, each participant completed a closing interview (which was also recorded) via Zoom. Participants answered questions about their experience in the study, including which version of the system they most preferred, any feedback or suggestions for system improvement, and their level of interest in using the system in the future. Interviews were transcribed automatically by Zoom, and qualitatively coded via inductive thematic analysis [12] in MaxQDA [24]. This study was approved by the University of Colorado Boulder's institutional review board (protocol number 23-0544).

4.3 Results

In this section, we present evaluation outcomes and an analysis of AI usage and errors.

4.3.1 Evaluation Outcomes. 18 participants completed three days of system usage, using a different interface on each day. Table 2 shows some of the basic statistics (averages) for each interface, including the number of events, the total time taken to complete the task, the time per event, the estimated time, estimated time per event, the perceived accuracy, the system usability score (SUS), and the average raw NASA-TLX score.

In general, participants rated all interfaces high on usability and on perceived accuracy. In addition, there was limited variation across the averages of the metrics for each interface, suggesting that participants were able to effectively use all three versions. The average number of final events is similar to (but slightly less than) those reported in other time use studies (roughly 17 [52] or 19 [23] per day). The cognitive load is slightly higher for the AI-only version, partially due to an increase in some participants' frustration with not being able to make manual modifications.

The confidence intervals of these metrics are generally overlapping across the interfaces. This indicates that in these metrics

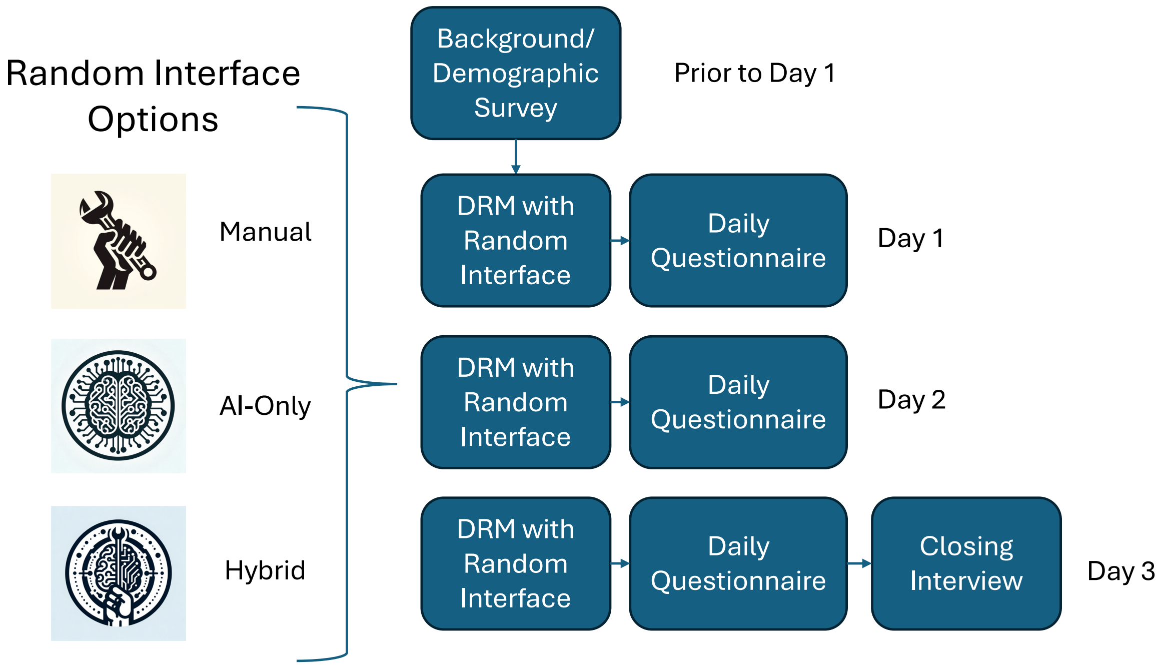

Figure 4: Specific tasks and ordering of the research study.

Metric	Manual	AI Only	Hybrid
Actual Duration (mins)	7.40	6.40	7.96
Estimated Duration (mins)	8.61	6.19	7.89
Number of Final Events	14.06	13.67	14.06
Accuracy Rating (1–5)	4.28	4.61	4.44
System Usability Score	85.42	82.50	85.69
Cognitive Load (TLX Raw)	32.93	34.16	31.87
Number of Manual Clicks	24.85	N/A	7.00

Table 2: Basic statistics about each interface (averages shown).

the interface either did not have a significant effect, or there was not enough data to show statistical significance. The one exception is the number of manual click events between the hybrid and manual interfaces. The 90% confidence interval for the click events for the manual interface is (17, 31) while the confidence interval for the hybrid interface is (3, 11). This provides clear evidence for the (unsurprising) finding that having access to the voice-interface reduced the need to make manual changes to the calendar state.

4.3.2 AI Usage and Error Analysis. Participants collectively used the AI voice interface 120 time, for both initial narratives and subsequent modifications. A dataset of narratives and spoken commands was created, including the transcription of the command, the returned event set from the LLM, hand-annotated intention(s) of the participant, and a "gold standard" event set. Across those 120 LLM calls (individual spoken commands transcribed), there were 412 requested changes made to the calendar state (as a single call could have multiple requested changes). The largest category of requested changes made to the calendar state is to create a new event (396 requests, 89.56% of all requests, see Table 3). This was followed by modifications to the calendar, such as deleting an event and moving an event's time (both had 8 requests, 1.94% of all requests). For example, P2 made the modification "so from 6:00 to 6:30 I was actually still sleeping" after their original narrative mentioned they woke up at 6AM. One request was to change the color of an event, which was not a feature provided by the calendar interface.

Category	Count	Percent
Create new event	369	89.56%
Delete event	8	1.94%
Move event to be earlier	7	1.70%
Change end time to lengthen event	5	1.21%
Split event into two	5	1.21%
Change end time to shorten event	4	0.97%
Replace old event with a new event	3	0.73%
Change start time to lengthen event	3	0.73%
Replace part of old event with new	2	0.49%
Change color of event (unsupported)	1	0.24%
Delete all gaps	1	0.24%
Keep event	1	0.24%
Modify event title	1	0.24%
Move event to be later	1	0.24%

Table 3: Categories of requested changes with counts and percentages.

Of the 412 requested changes, 341 of the changes were correctly implemented by the AI system. There were a total of 90 "failures" made by the system, defined as any action requested by the participant that was not completed by the system, or any action not requested by the participant that was completed by the system.

Category	Count	Percent
Wrong time	49	54.44%
Event creation failure	16	17.78%
Unexpected time change	9	10.00%
Extra event	8	8.89%
Mislabeled event	3	3.33%
Deleted wrong event	3	3.33%
Unexpected deletion	2	2.22%

Table 4: Categories of AI system failures with counts and percentages.

Table 4 shows the breakdown of the number of failures in each category. The most common failure was the system creating events at the wrong time. This error was compounded in narratives where there was a single "anchor time," and participants spoke about their events relative to that single anchor time. For instance, P12 explains part of their day as, "McDonald's probably took me 20 minutes to do that so after that head home by bus which took me around 30 minutes." Strings of event durations like these requires the LLM to perform temporal reasoning of various complexity, which often led to event time errors.

Some errors were related to inferences the LLM made about a participant's day. For example, P2 narrated "we went home and we had a very late dinner at 9:00 and then I went to, I worked a little bit more until about 10:30, 11:00 and that's when I slept." The resulting event set from the LLM included an unprompted "Sleep Preparation" event from 11:00 PM - 11:59 PM followed by a "Sleep" event starting at 11:59 PM. It is unclear why the LLM included this unprompted event. This could reveal an implicit belief of the LLM, that a logical human schedule requires a transition event between work and sleep, and that an hour was a probable amount of time for such preparation.

Another issue experienced by P1 was when trying to create overlapping events. P1 narrated "2:00 p.m. to 3:00 p.m. play with kittens" and then "2:30 to 3:00 p.m. chat on phone with friend", intending the phone call event to overlap with the second half of the kitten event. The LLM, however, shortened the kitten event from 2:00 - 2:30 rather than having the events overlap. P1 then made a voice correction saying "change play with kittens to be 2:00 to 3:00 and keep chat on phone with friend from 2:30 to 3:00." The LLM did create overlapping events as desired after this command.

4.4 User Experience and Subjective Feedback

Participants interacted with the system in a diversity of ways, particularly in how they utilized the voice interface of the calendar. The majority of participants (16, 89%) preferred the hybrid interface, while 2 participants (11%) preferred exclusive use of the manual interface without the AI features. This section contains a summary of the different themes identified from the think-aloud and interview transcripts related to the participant experience of using the calendar system.

4.4.1 Positive Impressions of AI Capabilities. Participants expressed that they liked to use the AI voice interface because of its speed (7 participants), ease of use (5 participants), accessibility (3 participants), and it being "more fun" to use (5 participants). Overall, participants seemed to be impressed with the capabilities of the AI, saying things like "Whoa! Whoa! Whoa! … I think it just made my life much easier" (P12) or "I'm really amazed by what I've actually seen today" (P11).

When reviewing the activities parsed by the the AI, P6 said, "Gap? I guess I must not have said anything from 9:30 to 10. Huh… Looking at my actual calendar, what happened at 9:30? Ha! [laughs] it's a gap. So it got it, very good, very good. So I'm impressed. This is better than I had any right to expect."

Participants particularly liked to use the voice interface to gather the bulk of the day's activities, and then make modifications with the manual interface. "I like having options. I think it's fun. The voice tool is fun. But I might not want it all the time, and then it's nice to be able to edit like what the voice tool did later, if you want to" (P3), and "I like the ability to be able to… edit, and improve, and fix, and drag and drop… after I've dictated everything. So I think the combination of the two things [modes of interaction]... gives me the agency to… make it what I want without having to… put in all the effort to… click and make every single event by hand" (P1).

4.4.2 Negative Impressions of AI Capabilities. Despite the general positive impressions of these AI capabilities, some participants experienced frustration and confusion while using the AI voice interface. Five participants expressed frustration that the system was slow (due to API latency). Five participants expressed uncertainty about the AI capability. Another five participants expressed confusion about what the AI did after making a change. P2 appeared to have a mental model that the AI could complete calendar modification tasks with similar ease to humans, and therefore had unmet expectations when the AI took a long time: "That took very long for just removing a gap. I was very clear in my instructions" (P2).

P7, when using the manual interface (after being exposed to the AI interface on the previous day), decided not to use the AI voice feature at all and only use manual interaction. "That's just how I do things, how I've always done it… It's what I'm used to, it's easier to type it… and it felt a little weird to like, tell the story of my day" (P7). Other participants felt that they could simply complete the task faster or more easily using the manual interface: "It's faster to just do it manually" (P3), "I prefer not to use the voice [interface]. Typing was much more easier for me" (P17). This was not true for all participants, as P14 explains, "There is a target demographic that would use it not only for accessibility purposes, which is very important… but also just maybe, like connecting it to another device [such as a home assistant]."

Two participants noted that the voice interaction provides less privacy, since other nearby individuals would be able to overhear the details of how they spend their time, which could be sensitive information. P17 said that with the manual interface, "I could decide to update a few things about my day and not want someone else to know what I'm doing. That's why I prefer typing… If I decide to use the AI, I would have to talk, and someone beside me could actually hear what I was doing."

Lastly, some participants expressed a desire to have more control over the AI system. Participants wanted to keep a tight loop between their own actions and the actions of the AI. P3 wanted to have the system act more like Apple's Siri digital assistant[9], where the system would prompt the user to confirm or reject a proposed change (such as a new calendar event). P7 mentioned that the lack of control over the system was one reason they did not want to use it: "If I was typing it, I knew it was accurate because I actually clicked on the box. So there was a bit more of a like guarantee I was in control" (P7).

4.4.3 Diverse Interactions with Calendar System. There was a spectrum of voice interaction styles ranging from short, simple statements to long, complex narratives with side comments. Participants acting most cautiously would provide short, simple commands to the system, asking the system to create single events at a time. For example, P1 would provide commands such as "9:30 a.m. to 10:30 a.m. making breakfast" and then stop the recording in order to see what the system would do. Other participants would share a long narrative about their day, with one participant speaking a narrative that contained 21 events. (The system was able to parse out 20 of them correctly.)

P4 began the study with short-form commands, saying, "'I just assumed that I needed to speak like a computer in order for it to be able to know, to put it in there [the calendar]." They thought the cognitive burden of using a voice interface was higher than clicking and dragging. They said it was largely due to the multiple steps of cognitive processing—they not only had to remember their day, but they also had to speak "to the system in a way that I thought the system would understand." This translation step added to the cognitive burden, and was (at least partly) the result of uncertainty about the capability of the AI system: "I probably would have spoken to it differently… if I believed that it could parse like full… natural language."

P1 eventually did start to use longer narratives in the calendar, and described the switch as "I was thinking of it more like a calendar at first… as a human creating that [the events], that's how I have to do it… event-by-event. But then, if the system just does everything for me… it was much easier for me to just like spew it all out" (P1).

P14 highlighted the need for improved training to understand the system's capability, saying "I think with better instructions, I think most people would know what they could say and what they couldn't say… more explicit rules on what it's good at and how the system breaks… I don't think most people would test that out on their own, so it'd be nice to let them know the types of things they could do." P4 had a similar experience, learning about the capabilities of the system through use: "Having learned that it would be more responsive to natural language… it didn't seem as taxing (P4)."

5 Discussion

5.1 Collecting Time-Use Data with a Voice Interface

The results of this study indicate opportunities and challenges for integrating an AI-powered voice interface into time-use studies. Despite the errors made by the system, all but two participants preferred to use the voice interface over manual data entry. Although there is generally no independent verification of accuracy of self-reported time-use studies, the reported level of perceived accuracy was relatively high for all 3 interfaces studied, and there was no significant difference in perceived accuracy ratings across the interfaces. The number of reported events (a common metric of data quality in time-use studies [35]) was also similar across all interfaces, providing evidence for the validity of using the voice interface. Given the high cost of conducting phone interviews to gather data (in the ATUS [55], for example), an AI-augmented voice system could be utilized to parse narrated transcripts and reduce operational costs of running time-use studies.

A voice interface, such as TellTime, could be integrated with existing electronic instruments for gathering time-use data, such as the Modular Online Time Use Survey (MOTUS) [52]. A study of the MOTUS system showed high attrition in participation between the participants previewing the diary and actually starting to use it. Giving participants the option of simply speaking about their day in order to begin using the data collection instrument may reduce the barrier to participation in time-use studies.

In TellTime, the system was often able to highlight gaps in the recorded data by creating gap events, implicitly prompting participants to fill out the missing time segment. While the present system did not explicitly ask the participant to fill out the missing time, it would be possible to create additional checks and verifications that highlight missing time on the calendar and explicitly prompt the participant to provide information. In addition, a chatbot could also be used to allow for two-way natural language dialog, with the chatbot playing the role of a traditional time-use surveyor conducting a phone interview, asking follow up questions about gaps of time or clarifications about additional context being captured. The chatbot could be integrated with the calendar interface, highlighting empty time segments in addition to prompting with natural language. This would follow a similar intention to previous work that utilized a chatbot to capture other kinds of self-reported personal informatics data [72].

A challenge in having an LLM parse a structured narrative is that the LLM would make inferences about the start and stop times of activities in the presence of an ambiguous narrative. After P4 spoke the narrative, "At 6am I woke up, and at 6:15 I got in the shower," the AI system made a showering event from 6:15–6:30 AM, inferring that the shower would take 15 minutes. Although some participants liked the inferences and saw them as a way to assist in reconstructing their day despite their own poor memory, this represents a risk to the integrity of fine-grained time-use surveys.

Model collapse happens when output from an AI system is used to train future iterations of the AI system, and results in the model forgetting the true underlying distribution of data (from the "ground truth," non–AI-generated data). After model collapse, models are less likely to predict long-tail (less likely) events [62]. In the context of time use, this might result in the AI inferring durations that are more standard across individuals, potentially resulting in a convergence of represented time use in the dataset. The risk of model collapse can be reduced by providing additional data (not generated by previous versions of the model) in subsequent training operations. Regardless, this issue—of participants relying on AI

[9] https://www.apple.com/siri/

inferences to record their time use—should be better understood and mitigated before using AI in any large-scale, real-world data collection study.

5.2 Design Considerations for Self-Tracking Time Use in Personal Informatics

The results contribute to the understanding of AI for accessibility in self-tracking and self-tracking via spoken narrative.

5.2.1 AI for Accessibility in Self-Tracking. Multiple participants noted that the voice interface can increase the accessibility of self-tracking, such as for the blind and visually impaired or those who have difficulty navigating traditional interfaces. The AI-augmented system seemed particularly helpful for participants who had trouble using the manual interface, either due to lack of ability with a mouse or trackpad, or not having prior experience with digital calendar systems. The voice interface was simply easier to use and required less effort for these participants. However, the most preferred version of the system was the hybrid interface. Participants liked to use the AI for a "one shot" bulk construction of the day and then to make manual modifications where the AI needed correction. Someone who is using the voice interface for accessibility reasons may not be able to engage in this form of hybrid interaction. This motivates further iteration on the design of the voice interface, particularly for making fast, precise modifications, which could be done using a more traditional multi-model NLP pipeline or a multi-step LLM pipeline.

5.2.2 Data Collection via Narrative. Using the AI-augmented voice interface enabled participants to collect time-use data via a spoken narrative. For some participants, the narrative process of recalling the day led to increased recall. For example, P12 said, "When I was talking to the AI, I kept on recalling what happened yesterday, and some minor details which, on a normal occasion, could be neglected by my mind suddenly just popped out." This finding suggests that the verbal narrative helps to "guide the recall process," a strategy of other technology-assisted recall methods [37]. It is possible that using a spoken narrative helps maintain a forward-seeking temporal recall process, where recalling one event of the day helps to cue the recall of subsequent events [33].

Previous work described semi-automated tracking as when systems gather data both automatically from sensors and from self-report [41]. Having the system parse structured data from a free-form spoken narrative is a different kind of semi-automated tracking. Here, however, the system serves as more of a "tracking assistant" that helps to properly structure the data spoken by the participants. The benefits are a reduced effort of manually inputting the data, and not requiring any external data or sensors.

The AI's generation of the events on the calendar required the participant to closely inspect the results of the calendar in order to implicitly "approve" or, by deleting or undoing, "reject" the changes made by the AI. This meant that the participants were closely engaged with a record of their time use. By having another agent (in this case, the AI) make its best guess at how the participant spent their time, it may actually increase participant engagement, as they have to closely look at what the AI is proposing and compare the results with their own memory. This suggests another potential approach of AI in personal informatics: the LLM creates an estimate of individuals' personal data, and that estimate serves as a boundary artifact that prompts those individuals to engage more deeply with their own experience, either through correction or confirmation. This can help increase engagement with time-tracking systems, a noted problem with more automated systems [11].

5.3 Human–AI Interaction on a Calendar

In a hybrid model of interaction, where the human and the AI collaborate on a shared artifact (in this case, a calendar), failures of the AI system differ in their effect on the outcome of the task. As noted in the results, the most common failure (54% of failures) was setting an incorrect time on an activity. This particular failure is potentially less troublesome than, say, mislabeling an event (3% of failures) when embedded in an interface with hybrid interaction. An incorrect event time can be quickly fixed by dragging and dropping an event to its proper place on the calendar, while renaming an event requires at least one click, followed by typing, in order to correct the mistake.

This interaction pattern matches the usage of most participants, who preferred to use the AI voice interface for recording the bulk of their daily events, eliminating the burden of typing out the details of all of the individual activities. The participants would then manually make changes to the suggested reconstruction created by the AI system. Part of this was due to AI system latency. The length of the prompt affected speed considerably—the longer the prompt, the higher the latency [57]. Given that the entire event list and spoken narrative is included in the prompt each time a call to the LLM was made in this calendar system, and the entire event list is requested as the output of the LLM, requests for small edits might take roughly the same amount of time as requests to parse long narratives. Given similar latency, participants were frustrated because the AI "should have" handled some requests more quickly than others. This highlights a challenge in integrating LLMs into collaborative tasks: latency of LLMs can shape the ways in which participants interact with the system, generally. Many participants found it was faster to simply make the manual modifications to the calendar rather than speak a command and wait for the LLM to process the changes.

5.3.1 Hybrid Interaction to Mitigate LLM and User Errors. In the most naturally spoken narratives, events tended to have a temporal dependency on each other. Participants would mention an "anchor time" at which they started their day, and then subsequent events would be described in relation to previous events. For example, P12 explained part of their day by saying, "[Eating at a restaurant] probably took me 20 minutes to do... so after that, head home by bus, which took me around 30 minutes." The "anchor time" in the narrative didn't appear until later in the day (multiple events later), and the LLM would have had to make backward temporal inferences to determine the start time of the restaurant event. The LLM did not do so, but rather started creating events at the anchor time, resulting in a domino effect where the timing of almost every subsequent event was wrong. The participant then had to adjust the start and end time of nearly every event that appeared on the calendar manually. This shows that LLMs still have some potential areas of improvement in temporal reasoning.

These limitations motivate design strategies for integrating temporal reasoning into intelligent user interfaces. While a simple user-training approach would be to instruct users to speak with absolute start and stop times (rather than relative times) in order to make the narrative structure more clear to the LLM, the specific errors introduced by LLMs could be mitigated through interface design. For example, the issue where multiple events are created with the correct durations but with incorrect shifted start times suggests that a "select and drag *multiple events*" could allow a user to quickly correct this error by selecting all the events and manually shifting them.

Another interaction problem was when participants spoke narratives with ambiguity, generally in the form of under-specified start and stop times (that could not be specifically determined even with perfect temporal reasoning). In these cases, the LLM could not be expected to correctly identify the start and stop times, but the current interface forced it to. The interface could include visual cues for communicating uncertainty in the spoken narrative. This could take the form of color-coded events scheduled using an inferred time, or the use of fuzzy or jagged borders around events with less certain start and end times.

The Haze calendar interface provides one example of these cues, as it allows users to RSVP to events with uncertainty based on the start and stop time [5]. Hazy coloring of event edges could be integrated into TellTime in cases where the LLM is uncertain as to the start time or stop time of an event based on ambiguity in the spoken narrative. This would serve as a clear visual feedback to the user that their narrative requires clarification. This could be accomplished by prompting the LLM to provide additional meta data such as a level of certainty about the specific start and stop time for each event. This same strategy could be used when the LLM infers events that the user did not explicitly mention, but are implied by the narrative, such as a "Sleep Preparation" event between "Work" and "Sleep."

5.3.2 Mental Models of AI Capability. Participants' mental models of the capabilities of the AI system appeared to affect how the participants interacted with the system. In particular, there were two general styles of interaction (perhaps, two ends along a spectrum): short, simple commands, intended to create one event at a time, and longer narratives containing multiple events. Multiple participants expressed that their voice interaction was due to their previous experience with a calendar. The participants had a mental model of calendar interaction being one event at a time, and therefore that is how they spoke to the system, instructing the system to do what *they* (the human) would do.

In one case, when the AI was perceived as not having capability to understand and parse complex narratives, participants spoke with simple commands. Other participants spoke long narratives freely to the system and seemed to have confidence (or at least openness) to the system being capable of parsing their narrative. This finding echoes similar results from Calendar.help, where participants were pleasantly surprised about the capabilities of the system [14]. Future AI-augmented calendars should include clear training to users, along with examples showing the system's capability, to help users adapt to the calendar's novel capabilities.

5.3.3 Design for AI Explainability and Control. Transparency of AI systems can help support user perceptions of trust and positive user experience [66]. Participants expressed a desire to have more visibility into the AI system—both what the system actually did and why the system did what it did. The system in the present study did not provide explanations for why the system did what it did, which likely contributed to the participants' uncertainty. In addition, as the AI system directly replaced the entire calendar state after processing the transcript, it was not always entirely clear what changes the AI system had introduced.

Adding animations for moved events or fade-in and fade-out animations for new or deleted events could help reduce participants' change blindness and more quickly perceive the changes made by the AI system. The system could also capture the portion of the transcript that provided evidence for the creation or modification of an event, helping participants to more clearly see why the system made the changes it did. Alternatively, as we did display the live transcript, links could be embedded in parts of the transcript after processing, which, when clicked on, highlight the particular event that was created because of that particular part of the text.

Another approach would be to prompt the LLM to include some estimate of uncertainty with each parsed event, and then visualize that uncertainty in the calendar interface, as described previously. Displaying uncertainty can help participants understand what the system is doing, although there is a trade-off between the added cognitive load of viewing more encoded information and the benefits of being able to notice and interpret the uncertainty [2].

5.4 Limitations and Future Work

The use of the DRM as the participant task likely resulted in learning effects. Recalling the specific episodes of one's day is not necessarily a common practice, and participants expressed that their recall improved as the study went on. While the randomization of the order at which interfaces were presented helps to mitigate these learning effects across participants, they may have had an effect on the participants' overall experience of the system. A future study with a longer deployment period could have participants complete the DRM every day for a week (or some other "burn-in" period) with a manual interface before switching over to experiencing the AI interface to better mitigate these learning effects.

Another limitation of the study is the variability in the time of day at which the study was conducted. The study sessions were scheduled primarily for participant convenience, but this resulted in participants having a different time lag from the activities they were seeking to reconstruct. This is a known concern in time-use studies. Fortunately, a prior study shows that the accuracy of the DRM completed for the prior day or the current day is relatively similar [48]. Future work can include studying how these kind of systems are used *in situ* over time and in different contexts. Two participants expressed a privacy concern that standers-by could overhear them speaking aloud their narrative to the system. It would be worth exploring the specific contexts in which individuals use each of the interaction modalities: manual, hybrid, and voice-only.

Multiple participants expressed interest in having a full calendar system that uses an AI voice interface. This could help individuals in planning out their day, and result in a time management planning (TMP) system [49] with improved interaction accessibility. Participants were particularly interested in using such a system to plan out their day, and having the system provide intelligent recommendations about specific activities to do based on the context of the current day or from previously recorded data. Future work can explore incorporating the voice interface into a more fully featured calendar prototype or personal informatics system that enables participants to gather data about their time use and reflect on their longitudinal data.

6 Conclusion

This research introduced TellTime, an AI-augmented calendar system for gathering time-use data. TellTime allows for human–AI collaboration for completing the Day Reconstruction Method (DRM), facilitating natural language narrative data capture via a voice interface to reduce the burden of self-reported time tracking. We designed and implemented three versions of the interface (manual, AI-only, and hybrid) and evaluated the system with 18 participants in a randomized controlled trial. We coded and summarized 120 instances of participants using the AI-augmented voice interface and determined categories of 412 intentions of the voice commands. We analyzed interview reports of participants' experience to understand how voice interfaces can be used for calendar systems, including those for reflective personal informatics systems and those for collecting time-use data in population-scale research surveys. We provided design recommendations for AI-augmented voice interfaces for calendar systems, and highlighted challenges and opportunities for using AI to gather time-use data via spoken narrative.

Acknowledgments

We would like to thank all the participants who generously shared their time and feedback for this study. We would also like to thank all the members of the TMI Lab for providing feedback, and Anusha Venkateswaran for exploring initial prompt engineering. This project was supported by the Institute of Cognitive Science (ICS) at the University of Colorado Boulder.

References

[1] Priya Andurkar, Akshat Pandey, and Pranjali Deshpande. 2021. Calendar Delegation with Alexa. *International Journal of Computer Applications* 174, 29 (April 2021), 18–21. https://doi.org/10.5120/ijca2021921213

[2] Stavros Antifakos, Adrian Schwaninger, and Bernt Schiele. 2004. Evaluating the Effects of Displaying Uncertainty in Context-Aware Applications. In *UbiComp 2004: Ubiquitous Computing*, Nigel Davies, Elizabeth D. Mynatt, and Itiro Siio (Eds.). Springer, Berlin, Heidelberg, 54–69. https://doi.org/10.1007/978-3-540-30119-6_4

[3] Bon Adriel Aseniero, Charles Perin, Wesley Willett, Anthony Tang, and Sheelagh Carpendale. 2020. Activity River: Visualizing Planned and Logged Personal Activities for Reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces (AVI '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3399715.3399921

[4] Pauline M. Berry, Melinda Gervasio, Bart Peintner, and Neil Yorke-Smith. 2011. PTIME: Personalized assistance for calendaring. *ACM Transactions on Intelligent Systems and Technology* 2, 4 (July 2011), 40:1–40:22. https://doi.org/10.1145/1989734.1989744

[5] Ryan David Bowler, Benjamin Bach, and Larissa Pschetz. 2022. Exploring Uncertainty in Digital Scheduling, and The Wider Implications of Unrepresented Temporalities in HCI. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–12. https://doi.org/10.1145/3491102.3502107

[6] John Brooke. 1996. SUS: A "quick and dirty" usability scale. In *Usability Evaluation in Industry*, Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). Taylor & Francis, Bristol, PA, USA, 207–212. https://doi.org/10.1201/9781498710411-35

[7] Andy Brown, Caroline Jay, and Simon Harper. 2010. Audio access to calendars. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*. ACM, Raleigh North Carolina, 1–10. https://doi.org/10.1145/1805986.1806028

[8] Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A. Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 121–132. https://doi.org/10.1145/2750858.2804266

[9] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. TimeBench: A Comprehensive Evaluation of Temporal Reasoning Abilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1204–1228. https://doi.org/10.18653/v1/2024.acl-long.66

[10] Brendan I. Cohn-Sheehy, Angelique I. Delarazan, Jordan E. Crivelli-Decker, Zachariah M. Reagh, Nidhi S. Mundada, Andrew P. Yonelinas, Jeffrey M. Zacks, and Charan Ranganath. 2022. Narratives bridge the divide between distant events in episodic memory. *Memory & Cognition* 50, 3 (April 2022), 478–494. https://doi.org/10.3758/s13421-021-01178-x

[11] Emily I. M. Collins, Anna L. Cox, Jon Bird, and Daniel Harrison. 2014. Social networking use and RescueTime: the issue of engagement. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 687–690. https://doi.org/10.1145/2638728.2641322

[12] Juliet Corbin and Anselm Strauss. 2014. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th ed.). SAGE Publications, Los Angeles, CA.

[13] Felicia Cordeiro, Elizabeth Bales, Erin Cherry, and James Fogarty. 2015. Rethinking the Mobile Food Journal: Exploring Opportunities for Lightweight Photo-Based Capture. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3207–3216. https://doi.org/10.1145/2702123.2702154

[14] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar.help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2382–2393. https://doi.org/10.1145/3025453.3025780

[15] Mihaly Csikszentmihalyi and Reed Larson. 2014. Validity and Reliability of the Experience-Sampling Method. In *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, Mihaly Csikszentmihalyi (Ed.). Springer Netherlands, Dordrecht, 35–54. https://doi.org/10.1007/978-94-017-9088-8_3

[16] Anke Dittmar and Laura Dardar. 2015. Personal ecologies of calendar artifacts. *Journal of Interaction Science* 3, 1 (Dec. 2015), 2. https://doi.org/10.1186/s40166-015-0007-x

[17] Eurostat. 2009. *Harmonised European Time Use Surveys: 2008 Guidelines.* Office for Official Publications of the European Union, Luxembourg.

[18] William Fajzel, Eric D. Galbraith, Christopher Barrington-Leigh, Jacques Charmes, Elena Frie, Ian Hatton, Priscilla Le Mézo, Ron Milo, Kelton Minor, Xinbei Wan, Veronica Xia, and Shirley Xu. 2023. The global human day. *Proceedings of the National Academy of Sciences* 120, 25 (June 2023), e2219564120. https://doi.org/10.1073/pnas.2219564120

[19] Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of Time: A Benchmark for Evaluating LLMs on Temporal Reasoning. arXiv:2406.09170 [cs.CL] https://arxiv.org/abs/2406.09170

[20] Kimberly Fisher, Jonathan Gershuny, and A Gauthier. 2012. Multinational time use study: user's guide and documentation. *Centre for Time Use Research, University of Oxford*, Retrieved from https://www.timeuse.org/sites/ctur/files/public/ctur_report/5715/mtus-user-guide-r5.pdf.

[21] United Nations Economic Commission for Europe. 2013. *Guidelines for harmonizing time-use surveys.* Technical Report. United Nations Economic Commission for Europe.

[22] Arindam Ghosh, Morena Danieli, and Giuseppe Riccardi. 2015. Annotation and prediction of stress and workload from physiological and inertial signals. In *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2015)*. IEEE, Piscataway, NJ, USA, 1621–1624. https://doi.org/10.1109/EMBC.2015.7318685 ISSN: 1558-4615.

[23] Ignace Glorieux and Joeri Minnen. 2009. How many days? A comparison of the quality of time-use data from 2-day and 7-day diaries. *International Journal of*

Time Use Research 6 (Nov. 2009), 314–327. https://doi.org/10.13085/eIJTUR.6.2.314-327

[24] VERBI GmbH. 2019. MAXQDA 2020. https://www.maxqda.com

[25] Rúben Gouveia and Evangelos Karapanos. 2013. Footprint tracker: supporting diary studies with lifelogging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2921–2930. https://doi.org/10.1145/2470654.2481405

[26] Simone Grassini. 2023. Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology* 14 (July 2023), 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

[27] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval* 8, 1 (2014), 1–125. https://doi.org/10.1561/1500000033

[28] Daniel S. Hamermesh, Harley Frazis, and Jay Stewart. 2005. Data Watch: The American Time Use Survey. *Journal of Economic Perspectives* 19, 1 (March 2005), 221–232. https://doi.org/10.1257/0895330053148029

[29] Sebastian Hammerl, Thomas Hermann, and Helge Ritter. 2012. Towards a semi-automatic personal digital diary: detecting daily activities from smartphone sensors. In *Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '12)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2413097.2413128

[30] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Human Mental Workload, Vol. 52. North-Holland, Amsterdam, The Netherlands, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[31] Michael Hoefer and Stephen Voida. 2023. Being, Having, Doing, and Interacting: A Personal Informatics Approach to Understanding Human Need Satisfaction in Everyday Life. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2593–2610. https://doi.org/10.1145/3563657.3596120

[32] Alexander Hoelzemann and Kristof Van Laerhoven. 2023. A Matter of Annotation: An Empirical Study on In Situ and Self-Recall Activity Annotations from Wearable Sensors. http://arxiv.org/abs/2305.08752 arXiv:2305.08752 [cs].

[33] Marc W. Howard and Michael J. Kahana. 2002. A Distributed Representation of Temporal Context. *Journal of Mathematical Psychology* 46, 3 (June 2002), 269–299. https://doi.org/10.1006/jmps.2001.1388

[34] Raghav Jain, Daivik Sojitra, Arkadeep Acharya, Sriparna Saha, Adam Jatowt, and Sandipan Dandapat. 2023. Do Language Models Have a Common Sense regarding Time? Revisiting Temporal Commonsense Reasoning in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6750–6774. https://doi.org/10.18653/v1/2023.emnlp-main.418

[35] F. Thomas Juster. 1986. Response Errors in the Measurement of Time Use. *J. Amer. Statist. Assoc.* 81, 394 (1986), 390–402. https://doi.org/10.2307/2289228

[36] Daniel Kahneman, Alan B. Krueger, David A. Schkade, Norbert Schwarz, and Arthur A. Stone. 2004. A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science* 306, 5702 (2004), 1776–1780. https://doi.org/10.1126/science.1103572

[37] Evangelos Karapanos. 2020. Technology-assisted reconstruction: a new alternative to the experience sampling method. *Behaviour & Information Technology* 39, 7 (July 2020), 722–740. https://doi.org/10.1080/0144929X.2019.1608303

[38] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A. Kientz. 2012. Lullaby: A capture & access system for understanding the sleep environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 226–234. https://doi.org/10.1145/2370216.2370253

[39] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–21. https://doi.org/10.1145/3491102.3517457

[40] Young-Ho Kim, Jae Ho Jeon, Eun Kyoung Choe, Bongshin Lee, KwonHyun Kim, and Jinwook Seo. 2016. TimeAware: Leveraging Framing Effects to Enhance Personal Productivity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/2858036.2858428

[41] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 67:1–67:28. https://doi.org/10.1145/3130930

[42] Young-Ho Kim, Sungdong Kim, Minsuk Chang, and Sang-Woo Lee. 2022. Leveraging Pre-Trained Language Models to Streamline Natural Language Interaction for Self-Tracking. In *Proceedings of the 2nd Workshop on Bridging Human-Computer Interaction and Natural Language Processing (NAACL '22)*. Association for Computational Linguistics, Kerrville, TX, USA.

[43] Amel Ksibi, Ala Saleh D. Alluhaidan, Amina Salhi, and Sahar A. El-Rahman. 2021. Overview of Lifelogging: Current Challenges and Advances. *IEEE Access* 9 (2021), 62630–62641. https://doi.org/10.1109/ACCESS.2021.3073469

[44] Clayton Lewis. 1982. *Using the "Thinking-Aloud" Method in Cognitive Interface Design.* Technical Report RC 9265. IBM T. J. Watson Research Center, Yorktown Heights, NY.

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

[46] James J. Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B. Strub. 2006. Fish'n'Steps: encouraging physical activity with an interactive computer game. In *Proceedings of the 8th international conference on Ubiquitous Computing (UbiComp'06)*. Springer-Verlag, Berlin, Heidelberg, 261–278. https://doi.org/10.1007/11853565_16

[47] Jonathan Liono, Flora D. Salim, Niels Van Berkel, Vassilis Kostakos, and A. K. Qin. 2019. Improving Experience Sampling with Multi-view User-driven Annotation Prediction. In *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. IEEE, Kyoto, Japan, 1–11. https://doi.org/10.1109/PERCOM.2019.8767394

[48] Kai Ludwigs, Lena Henning, and Lidia R. Arends. 2019. Using the Day Reconstruction Method - Same Results when Used at the End of the Day or on the Next Day? *International Journal of Community Well-Being* 2, 1 (May 2019), 61–73. https://doi.org/10.1007/s42413-019-00017-x

[49] John R. Lund and Jason Wiese. 2021. Less is More: Exploring Support for Time Management Planning. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 392–405. https://doi.org/10.1145/3461778.3462133

[50] Lena Mamykina, Daniel A. Epstein, Predrag Klasnja, Donna Sprujt-Metz, Jochen Meyer, Mary Czerwinski, Tim Althoff, Eun Kyoung Choe, Munmun De Choudhury, and Brian Lim. 2022. Grand Challenges for Personal Informatics and AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. ACM, New Orleans LA USA, 1–6. https://doi.org/10.1145/3491101.3503718

[51] Talie Massachi, Grant Fong, Varun Mathur, Sachin R. Pendse, Gabriela Hoefer, Jessica J. Fu, Chong Wang, Nikita Ramoji, Nicole R. Nugent, Megan L. Ranney, Daniel P. Dickstein, Michael F. Armey, Ellie Pavlick, and Jeff Huang. 2020. Sochiatrist: Signals of Affect in Messaging Data. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–25. https://doi.org/10.1145/3415182

[52] Joeri Minnen, Ignace Glorieux, Theun-Pieter Van Tienoven, Sarah Daniels, Djiwo Weenas, Jef Deyaert, Sarah Van Den Bogaert, and Sven Rymenants. 2014. Modular Online Time Use Survey (MOTUS) – Translating an existing method in the 21st century. *Electronic International Journal of Time Use Research* 1, 11 (2014), 73–93.

[53] Pragnesh Jay Modi, Manuela Veloso, Stephen F. Smith, and Jean Oh. 2005. CMRadar: A Personal Assistant Agent for Calendar Management. In *Agent-Oriented Information Systems II*, Paolo Bresciani, Paolo Giorgini, Brian Henderson-Sellers, Graham Low, and Michael Winikoff (Eds.). Springer, Berlin, Heidelberg, 169–181. https://doi.org/10.1007/11426714_12

[54] Evangelos Niforatos and Evangelos Karapanos. 2015. EmoSnaps: a mobile application for emotion recall from facial expressions. *Personal and Ubiquitous Computing* 19, 2 (Feb. 2015), 425–444. https://doi.org/10.1007/s00779-014-0777-0

[55] Bureau of Labor Statistics. 2023. American Time Use Survey User's Guide.

[56] Leysia Palen. 1999. Social, individual and technological issues for groupware calendar systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 17–24. https://doi.org/10.1145/302979.302982

[57] Proxet. 2024. Understanding Latency in LLM: The Impact of Token Generation on Response Time. https://www.proxet.com/blog/llm-has-a-performance-problem-inherent-to-its-architecture-latency

[58] Elizabeth Quinlan. 2008. Conspicuous Invisibility: Shadowing as a Data Collection Strategy. *Qualitative Inquiry* 14, 8 (Dec. 2008), 1480–1499. https://doi.org/10.1177/1077800408318318

[59] L.D. Rosen, K. Whaling, L.M. Carrier, N.A. Cheever, and J. Rokkum. 2013. The Media and Technology Usage and Attitudes Scale: An empirical investigation. *Computers in human behavior* 29, 6 (Nov. 2013), 2501–2511. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338964/

[60] Pedro Sanches, Kristina Höök, Elsa Vaara, Claus Weymann, Markus Bylund, Pedro Ferreira, Nathalie Peira, and Marie Sjölinder. 2010. Mind the Body! Designing a Mobile Stress Management Application Encouraging Personal Reflection. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)*. ACM Press, New York, NY, USA, 47–56. https://doi.org/10.1145/1858171.1858182

[61] Adam Shaw. 2023. FullCalendar.

[62] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. The Curse of Recursion: Training on Generated Data Makes Models Forget. https://doi.org/10.48550/arXiv.2305.17493

[63] Jay Stewart. 2018. Getting Started with Time Use Data. *Journal of Time Use Research* 14, 1 (Dec. 2018), 24–33. https://doi.org/10.32797/jtur-2018-2

[64] Zhaochen Su, Jun Zhang, Tong Zhu, Xiaoye Qu, Juntao Li, Min zhang, and Yu Cheng. 2024. Timo: Towards Better Temporal Reasoning for Language Models. *First Conference on Language Modeling*, Retrieved from https://openreview.net/forum?id=3TzGD95Jw1.

[65] O. Sullivan, J. Gershuny, A. Sevilla, P. Walthery, and M. Vega-Rapun. 2020. Time use diary design for our times - an overview, presenting a Click-and-Drag Diary Instrument (CaDDI) for online application. *Journal of Time Use Research* 15, 1 (Sept. 2020), 1–17. https://doi.org/10.32797/jtur-2020-1

[66] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 74–88. https://doi.org/10.1093/jcmc/zmz026

[67] United Nations (Ed.). 2005. *Guide to producing statistics on time use: measuring paid and unpaid work.* United Nations, New York.

[68] Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, and James Pustejovsky. 2014. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. http://arxiv.org/abs/1206.5333

[69] Yonatan Vaizman, Katherine Ellis, Gert Lanckriet, and Nadir Weibel. 2018. ExtraSensory App: Data Collection In-the-Wild with Rich User Interface to Self-Report Behavior. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3174128

[70] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The TempEval challenge: identifying temporal relations in text. *Language Resources and Evaluation* 43, 2 (June 2009), 161–179. https://doi.org/10.1007/s10579-009-9086-z

[71] Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking Temporal Reasoning for Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6389–6415. https://doi.org/10.18653/v1/2024.findings-acl.382

[72] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 87 (apr 2024), 35 pages. https://doi.org/10.1145/3637364

[73] Steve Whittaker, Vaiva Kalnikaite, Victoria Hollis, and Andrew Guydish. 2016. 'Don't Waste My Time': Use of Time Information Improves Focus. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1729–1738. https://doi.org/10.1145/2858036.2858193

[74] Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large Language Models Can Learn Temporal Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 10452–10470. https://doi.org/10.18653/v1/2024.acl-long.563

[75] Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the Future: Towards Explainable Temporal Reasoning with Large Language Models. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 1963–1974. https://doi.org/10.1145/3589334.3645376

[76] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704

A LLM Prompts

The following prompt was used when no events were on the calendar when the voice interface was used. These prompts were supplied as a "system" message to the LLM, as the first message in a conversation. The second message contained the transcript from the spoken narrative, and was supplied as a "user" message.

> You are to help parse the narrative the user provides into a structured list of activities that the user has done in their day.
> Please parse this narrative of my day into a csv list of activities like so:
> start time 1, end time 1, activity 1
> start time 2, end time 2, activity 2
> For any gap in time, create a GAP event. Only reply with the structured list, no other text or narration. Return all times in military time.

The following prompt was used when there were previously existing events on the calendar when the voice interface was used:

> You are to make modifications to the list of activities below, to comply with the changes requested by the user.
>
> When replying, respond with a complete list of activities in the same format as the original list, but with the requested changes.
> Include all activities. Replace "original" with "modified" if you modified the activity, or add "new" if you created a new activity. All times are in military time.
> If no changes are requested, reply with the original set of activities, no other text or narration.
> The list is formatted as follows:
> id 1, activity 1, start time 1, end time 1, original
> id 2, activity 2, start time 2, end time 2, original
>
> Only reply with the structured list, no other text or narration. Return all times in military time.

When events were already on the calendar and the voice interface was used, the previous events were appended to the transcript in the following format:
id 1, activity 1, start time 1, end time 1, original
id 2, activity 2, start time 2, end time 2, original
...

B Participant Instructions

The following instructions were provided, one for each interface.

B.1 Instructions for Manual Interface

> This is a basic calendar interface, similar to Google calendar or Microsoft outlook. You can create and edit events by clicking on the calendar and dragging the mouse. You can edit events by clicking on them, or clicking and dragging. Here, you will reflect on your day, and enter all the activities that you engaged in either today or the previous day. Please break up your day into episodes and create a calendar event for each episode. Make sure your activities cover the entire time from midnight to midnight. Click "Finish" when you are done.

B.2 Instructions for AI-only Interface

> This is a basic calendar interface, similar to Google calendar or Microsoft outlook. However, instead of using the mouse to create events, you will be speaking a narrative of your day to the system, which will then create the calendar events for you. You can click the "Start Recording" button and just explain what you did yesterday, and click "Stop Recording" when finished. If you need to make changes to the events, you can

click “Start Recording” again, and instruct the system to make changes for you. You can undo the last change made by the system by clicking “Undo”. You cannot interact with the calendar directly, you must use the voice interface. Make sure your activities cover the entire time from midnight to midnight. Click “Finish” when you are done.

B.3 Instructions for Hybrid Interface

This is a basic calendar interface, similar to Google calendar or Microsoft outlook. However, instead of only using the mouse to create events, you can speak a narrative of your day to the system, which will then create the calendar events for you. You can click the “Start Recording” button and just explain what you did yesterday, and click “Stop Recording” when finished. You can also ask the system to make modifications to existing events (by clicking “Start Recording” and speaking your desired changes). You can undo the last change made by the system by clicking “Undo”. You can also use the mouse to create and edit events by clicking and dragging. Make sure your activities cover the entire time from midnight to midnight. Click “Finish” when you are done.