Int. J. Human–Computer Studies 215 (2026) 103878

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

From prediction to explanation: Using screen text to understand smartphone use and user behaviour

Songyan Teng*, Hong Jia, Simon D'Alfonso, Vassilis Kostakos

The University of Melbourne, Parkville, Melbourne, 3010, Victoria, Australia

HIGHLIGHTS

- Leveraged smartphone screen text to predict and explain user behaviours.
- Demonstrated interpretability of fine-tuned large language models (LLMs).
- Linked digital screen content with offline, real-world user activities.
- Developed a framework for explainable smartphone sensing research.

ARTICLE INFO

Keywords:
Screen text
Large language models
Smartphone sensing
User behaviour
Behavioural explanation

ABSTRACT

Smartphones are essential to daily life, and their rich data streams have been used to study how people use their phones, and more broadly human behaviour. While previous research has largely focused on app usage and keystroke dynamics to predict smartphone use, these analyses are typically limited to making predictions rather than providing explanations or reasoning for observed behaviours. In this exploratory study, we investigate the potential of leveraging screen text and large language models (LLMs) to uncover insights and reasoning about user behaviour. Using a dataset of over 100 million on-screen words collected from 21 participants over two weeks, we explore multiple ways to use screen text and LLMs for three tasks: predicting the next app a user will open, inferring what real-world activities they are engaged in, and understanding how they interact within apps. Additionally, we demonstrate the interpretive capabilities of LLMs, highlighting their potential to explain the reasoning behind observed user actions. Our findings suggest that screen text holds promise for providing deeper insights into both digital and real-world human behaviour. We discuss the broader implications of our findings, including enhancing user experience and enabling privacy-preserving, on-device analysis, while proposing future research directions in screen text analysis.

1. Introduction

Smartphones have transformed the way individuals interact with technology, serving not only as communication tools but also hosting sensors capable of continuously capturing a range of data streams (Masoud et al., 2019), such as app usage (Ferreira et al., 2015) and physical location (Huang et al., 2016). These data streams have enabled new methods for understanding human behaviour with greater precision and in real-time, and these opportunities have been studied across various fields such as health (Beiwinkel et al., 2016), psychology (Wahle et al., 2016), and education (Zhang et al., 2015). Traditional approaches to studying human behaviour, such as surveys and interviews, often lack contextual richness (Pejovic et al., 2015). In contrast, smartphone sensing offers a passive and unobtrusive method to capture a continuous record of an individual's behaviour (Trifan et al., 2019), which can be analysed to infer behavioural patterns (Tonti et al., 2021).

Yet, the textual content users view on their screens remains underexplored. While prior studies have analysed app usage (Zhao et al., 2019a) and keystrokes (Ciman et al., 2015), on-screen text offers richer behavioural insight (Ram et al., 2019). It can reveal the context behind actions (Teng et al., 2024b), such as reasons for app switching or how digital interactions relate to offline environments. Understanding not just *what* users do but *why* they do it is essential for deeper behavioural

* Corresponding author.
Email addresses: songyan.teng@student.unimelb.edu.au (S. Teng), hong.jia@unimelb.edu.au (H. Jia), dalfonso@unimelb.edu.au (S. D'Alfonso), vassilis.kostakos@unimelb.edu.au (V. Kostakos).

https://doi.org/10.1016/j.ijhcs.2026.103878
Received 23 July 2025; Received in revised form 9 June 2026; Accepted 13 June 2026
Available online 15 June 2026

insight. Given individual variability, we also investigate whether fine-tuned large language models (LLMs) can enhance these insights by leveraging the semantics of screen text in predictive models.

This exploratory study examines how LLMs can be used not just to predict behaviour from screen text (Teng et al., 2024a), but also to generate interpretable explanations. Rather than using LLMs solely for prediction, we explore their potential to bridge the gap between digital behaviour and underlying motivation. Since screen text is inherently semantic and context-rich, LLMs are uniquely positioned to extract behavioural meaning from this data, in contrast to traditional machine learning approaches often used with sensor-based data. This shift from prediction to explanation opens the door to more transparent, personalised, and user-aligned systems.

We demonstrate this through three experiments: (1) predicting the next app a user will open; (2) linking screen text with user-reported real-world activities using experience sampling method (ESM) data; and (3) identifying the in-app activities users perform. These scenarios allow us to assess both the potential and the challenges of applying fine-tuned LLMs to smartphone data, while highlighting trade-offs between general and personalised models and the ethical considerations of working with sensitive behavioural data.

We conduct our experiments using a smartphone sensing dataset of over 100 million on-screen words collected from 21 participants over a two-week period. This dataset captures the textual content displayed on participants' smartphone screens, known as screen text. ESM responses are also collected from participants, which provide self-reported data on what activities they were engaged in at specific moments throughout the day.

The first experiment investigates next-app prediction, where anticipating a user's next app can improve experience design (Guo et al., 2019) and resource allocation (Yan et al., 2012). We evaluate whether fine-tuned LLMs can use screen text to identify behavioural context and generate explanations of app-switching decisions. We fine-tune Meta's Llama 3-8b-Instruct model (Meta, 2024a) on screen text paired with subsequent app usage to evaluate how personal context affects performance and privacy.

The second experiment links screen text with self-reported activity data from ESM (van Berkel et al., 2017). We fine-tune both individualised models (trained on a user's own data) and generalised models (trained on others' data) to assess their ability to predict offline activities (de Vries et al., 2020). We also evaluate how prompting the model to generate explanations can improve our understanding of the relationship between digital behaviour and real-world activity.

The third experiment explores in-app activity categorisation. Unlike app-level tracking, understanding specific in-app behaviours is more complex (Cao and Lin, 2017). For example, within a health app, one user may focus on fitness plans while another reads about nutrition (Goodyear et al., 2018). We use screen text to summarise and categorise such activities, showing how LLMs can abstract behavioural meaning and support privacy-preserving summarisation.

The contributions of our paper are threefold:

- We present an exploratory framework for using LLM-based methods to analyse screen text data, focusing on LLM fine-tuning and its associated trade-offs.
- We demonstrate how generating explanations for predictions can provide deeper insights into why certain behaviours occur, exploring an important gap in interpretability within previous studies.
- We develop experiments and preliminary methods that can guide future research in creating fine-tuned, privacy-preserving LLMs for user behaviour studies.

2. Related work

2.1. Smartphone sensing for understanding behaviour

Smartphone sensing has been used as a powerful tool for understanding human behaviour, leveraging the ubiquity of smartphones and their capacity to collect real-time data (Harari et al., 2016). Sensors such as GPS, accelerometers, and app usage logs have been used to infer behavioural traits including social interaction (Fulford et al., 2021), mental health (Vega et al., 2022), and stress (Zhang et al., 2018, 2024a). This passive, unobtrusive data collection offers high-granularity, context-rich insights, avoiding limitations of traditional methods like surveys, which can suffer from recall bias and participant burden (Jager et al., 2020). Such sensing is especially useful in fields like health, education, and psychology, where nuanced behavioural understanding is essential.

While many smartphone sensing approaches exist, the specific ways people use their phones and the motivations behind their actions remain underexplored. Prior work has used smartphone text data, such as messages (Stachl et al., 2020) and social media interactions (Karim et al., 2020), to examine user behaviour, while keystroke logging has revealed links to emotion and stress (Ferreira et al., 2015; Shapsough et al., 2016; Sağbaş et al., 2020). However, these methods do not capture the full scope of on-screen content and often miss key contextual cues. High-frequency screen capture, or Screenomes (Reeves et al., 2019), provides a richer view but is computationally intensive and difficult to annotate at scale. The screen text sensor (Teng et al., 2024a), which uses Android's accessibility services to collect textual screen content, offers a lightweight and scalable alternative. However, techniques for extracting deep behavioural insights from this data are still limited. Our work addresses this gap by developing methods for analysing screen text, enabling explanations of both how people use their smartphones and why they engage in specific behaviours. Through this, we aim to demonstrate the utility of screen text for behavioural insight and explanatory analysis.

2.1.1. Smartphone next app prediction

Next-app prediction is a well-established research area in behaviour computing. Anticipating the app a user will open next can streamline interaction, improve battery life through resource optimisation, and enable more personalised services. Traditional approaches rely on machine learning models using features like time of day (Sarker and Salah, 2019), location (Solomon et al., 2022), and app usage history (Khaokaew et al., 2024). Much of this work has focused on maximising predictive accuracy. For example, Yan et al. (2012)'s FALCON system used a cost-benefit classifier on spatiotemporal data to pre-load likely apps, while Katsarou et al. (2022) applied LSTM models to recent app sequences. Markov models and statistical predictors for next-app prediction have also incorporated cues such as last used app or current location (Parate et al., 2013). Middleware platforms like FutureWare extend these capabilities to anticipatory mobile computing (Mehrotra et al., 2021). However, these methods largely treat next-app prediction as a purely algorithmic task, prioritising accuracy over interpretability. Such models lack the ability to explain predictions, making it difficult to understand which user behaviours contributed to the outcome. Screen text offers a richer behavioural context by capturing the content users interact with (Fang et al., 2024), allowing us to infer users' focus and reasoning. LLMs have the potential to provide natural-language explanations for predictions, shifting emphasis from pure accuracy to transparency and interpretability. The goal is to complement established predictors with transparent, context-aware rationales for each prediction, rather than to compete on accuracy.

2.1.2. ESM real-world activity prediction

Experience Sampling Method (ESM) is widely used to collect in-situ, real-time data on individuals' behaviours and experiences (Trull and Ebner-Priemer, 2009). ESM questionnaires enable users to self-report real-world activities, including those not directly observable through smartphone data, and these can be analysed alongside smartphone usage patterns (van Berkel et al., 2017). Traditional approaches rely on sensor data such as GPS, accelerometers, and app usage to infer activities (Beames et al., 2024), offering valuable context over time. However, the integration of screen text presents new opportunities to improve model explainability. Since users frequently engage with their phones, the content on their screens often mirrors or influences their offline behaviour (Melumad and Pham, 2020). For instance, viewing travel websites may indicate upcoming trips, while reading recipes might precede grocery shopping. Analysing this screen text can offer deeper insight into the connection between digital interactions and real-world activity, as well as the smartphone usage sequences that precede them (Zhang et al., 2024b).

2.1.3. App usage activity

Traditional analyses of app usage focus on frequency and duration (Ferreira et al., 2014), app transitions (Monge Roffarello and De Russis, 2022), and app types (Böhmer et al., 2011), offering insight into general behavioural patterns. For instance, frequent messaging may reflect social activeness (Twenge et al., 2019), while high camera use may indicate openness (Stachl et al., 2019). However, these methods reveal only which apps are used, not how they are used. This is an important distinction as individual intentions within apps can vary widely. Screen text analysis can close this gap by uncovering the content users engage with inside apps (Teng et al., 2024a). For example, social media text may show whether a user is passively browsing posts or actively conversing. This richer understanding enables more precise inferences about intentions, emotional states, and behavioural patterns.

2.2. LLMs in behavioural analysis

Large language models (LLMs) are increasingly used in behavioural analysis due to their ability to process complex, unstructured text data. Models like GPT-4 (OpenAI, 2023) and Llama 3 (Meta, 2024a) have been applied to sentiment analysis (Teng et al., 2024b), affect detection (Zhang et al., 2024c), and behavioural trend prediction.

A key strength of LLMs lies in their capacity to interpret unstructured and context-rich data (Cheung, 2024), which is common in smartphone interactions. Traditional approaches often depend on structured features and predefined categories (Moustafa et al., 2018), overlooking important behavioural context (Rahma and Wantini, 2024). In contrast, LLMs can recognise linguistic cues that signal emotions, intentions, or mental states (Li et al., 2023), enabling more nuanced behavioural interpretations. LLMs also support personalisation and adaptive interventions (Goslen et al., 2024), tailoring predictions to individual behaviour patterns (Ke et al., 2024). This is especially valuable in health (Cornet and Holden, 2018) and education (Kucirkova et al., 2021), where understanding individual differences is essential. Fine-tuning LLMs on behavioural datasets further enhances their relevance by capturing domain-specific patterns (Parthasarathy et al., 2024; Ferrara, 2024).

However, interpretability remains a core challenge (Brown, 2024; Singh et al., 2024). The opacity of LLMs makes it difficult to trace how outputs are generated. Explainability techniques such as chain-of-thought prompting (Wei et al., 2022) offer potential solutions but have yet to be fully explored in noisy, high-volume behavioural contexts like screen text. As LLMs continue to scale, balancing predictive performance, privacy, and interpretability will be critical for their responsible application in behavioural research.

3. Methodology

3.1. User study and dataset

We collected an in-the-wild dataset based on the methodology presented by Teng et al. (2024a), capturing data from 21 participants during a two-week field study. This dataset includes all textual data displayed on participants' smartphone screens, representing their smartphone interactions from a range of sources such as web browsing and media consumption. Textual data was collected using the screen text sensor, which is part of the AWARE-Light (van Berkel et al., 2022) smartphone sensor platform.

Experience sampling method (ESM) responses were also collected to capture participants' behaviours, thoughts, and feelings during their daily activities, providing insights beyond their smartphone usage (van Berkel et al., 2017). Participants were asked to describe their primary activity for the five minutes before receiving each questionnaire. They had to respond in a single sentence, and did so five times a day. This five-minute window was selected to capture a recent snapshot of user activity while minimising the burden on participants in recalling their activities (Teng et al., 2024a).

Participants were recruited by completing an expression of interest form and providing their consent. They were then given instructions for setting up the app, joining the study, and completing ESM questionnaires. Participants were also informed about the functionalities of the AWARE-Light application and briefed on what data would be captured and how they would have control over this data. A three-day testing period ensured smartphone compatibility with AWARE-Light and accurate data capture with the screen text sensor. Participants completed ESM questionnaires five times daily, between 10 a.m. and 6 p.m., delivered via notifications that expired after 15min (Ferreira et al., 2014). Participants were reimbursed upon completion of the study after the two-week period.

The study was approved by the University of Melbourne's Office of Research Ethics and Integrity.

3.2. Data preprocessing

The screen text sensor collects data in plain text, storing each text element and its screen coordinates as part of a single delimited string. We use these coordinates to arrange the text elements in a sequence that reflects their actual layout on the screen, following a top-to-bottom and left-to-right pattern. Since the screen text sensor captures all screen information whenever a change is detected, it frequently records overlapping text elements when users scroll. To reduce redundancy, we implement a de-duplication algorithm to remove these overlaps between successive screens, described in Appendix A. We then store this de-duplicated data in a JSON format.

To ensure that the input data matches our model's requirements, we construct a pipeline that loads and configures the base model and tokenizer for fine-tuning. The tokenizer is configured with a custom chat template that defines the conversation structure of the model, which can be adapted based on the chosen base model. Additionally, we implement a mapping strategy to distinctly assign conversation roles and content attributes, distinguishing between human and model-generated messages, which, in our experiments, correspond to the input prompt and expected output, respectively. Unlike traditional machine learning approaches, LLMs can process texts provided in their natural language structure (Mandvikar, 2023), without the need for preprocessing techniques such as manually creating vector representations of textual data.

Raw screen text inevitably contains noisy elements that are not directly indicative of user behaviour, including persistent interface labels and background text that remain on screen across multiple captures. We deliberately retained these elements in the input rather than filtering them out, for two reasons. First, our de-duplication algorithm already removes repeated text from overlapping screen captures during scrolling, so persistent UI elements are collapsed into a single occurrence per screen rather than amplified across many, which is the most common form of redundancy. Second, because LLMs are trained on large volumes of web and application text, they are generally tolerant of incidental UI noise and can extract meaning from the more semantically informative portions of a screen without explicit filtering. Aggressive rule-based removal of UI labels or advertisements would risk discarding genuinely informative content, so we treated noise tolerance as a property the model was expected to provide.

3.3. Analysis and experimental setup

In line with our study's goal of moving beyond accuracy, our use of LLMs focuses not only on generating predictions, but also on producing natural language explanations that reflect the model's reasoning. These explanations allow us to interpret model behaviour, understand its mistakes, and surface latent behavioural patterns that might otherwise be missed in purely quantitative evaluation. This dual use of LLMs as both predictors and explainers supports a richer understanding of smartphone use through screen text, and helps to address limitations in interpretability that often affect traditional machine learning models.

In our study, we employ Meta's Llama 3-8b-Instruct LLM as the baseline for all our experiments. The open-source nature of this model allows for greater control over fine-tuning and facilitates an exploratory approach to personalisation and context-sensitive tasks using screen text. While its relatively smaller size reduces computational costs, it still offers sufficient capacity for exploring nuanced patterns within screen text data. We utilise the Llama 3 tokenizer for processing and encoding the text data into a suitable format. We further customise the tokenizer by configuring it with Unsloth's Llama-3 chat template, a framework that enhances the model's ability to handle conversational data (Han and Han, 2024).

The fine-tuning process is conducted using a local copy of the Llama 3-8b-Instruct model obtained from the Hugging Face platform (Meta, 2024b). We use a batch size of 2 with gradient accumulation over 4 steps, resulting in an effective batch size of 8, to optimise memory usage. We also incorporate a learning rate of 2e-4 with linear decay and the AdamW optimizer (Loshchilov and Hutter, 2017) for more precise weight adjustments. Each fine-tuning cycle runs for 5 epochs to balance robustness and computational efficiency. All experiments are conducted on an NVIDIA 80GB A100 GPU with mixed precision training.

4. Experiments

4.1. Experiment 1: next-app prediction

The ability to predict the next smartphone application a user is likely to open can promote advancements in user interface design and enhance user experience. Additionally, next-app prediction can lead to more efficient resource management on devices, optimising battery life and processing power by pre-loading frequently used applications in advance. This experiment involves predicting the next app that a user will open, based on the screen text currently visible on their device and the app they are currently using. We focus on exploring how fine-tuned models could potentially harness nuanced screen text context to refine next-app predictions and provide explanations for app usage behaviours. In contrast with generalised approaches, which aggregate multiple users' data, we employ only individualised models for each participant in this experiment. Since each participant used a different set of apps, their data lacks the shared foundation that would ideally support a population-wide model. Therefore, training individualised models allows us to capture each participant's unique usage patterns more effectively.

4.1.1. Training method and data preparation

To accommodate the unique behaviours of each individual, we design a pipeline to train multiple fine-tuned next-app prediction models for each participant in our study. Additionally, we aim to identify the most effective training and evaluation window within the 14-day period for screen text data, achieving an optimal balance between prediction accuracy and the necessary amount of training data. We evaluate six different training-evaluation splits based on consecutive, chronologically-ordered time periods, as outlined in Table 1:

Table 1
Next-App prediction training and evaluation Splits.

Training Period	Evaluation Period	Notes
1 day	13 days	–
2 days	12 days	–
3 days	11 days	20/80 split
7 days	7 days	50/50 split
11 days	3 days	80/20 split
13 days	1 day	Leave-One-Out

We begin by loading the entire dataset and filtering it to include only the data from the participant currently being trained on. The data is then divided into training and evaluation sets according to the current split. Since the screen text sensor also records the app where the text was captured and stores this information, each screen text entry can be paired with the subsequent app to create screen text-next app pairs. The "next app" could be the same app paired with a different screen text entry, as we want our prediction to also capture when the user will stay on their current app. The final entry for each participant is discarded because there is no recorded next app for it. These pairs are then formatted into prompts for training, as detailed in Section 4.1.2. Altogether, the dataset comprises a total of 394 apps across a diverse range of categories including social media, internet, and shopping.

4.1.2. Prompt design

Our goal is to fine-tune the base model using (screen text) and (next-app) pairs to help it understand the associations between the text displayed on the screen and the user's subsequent app choices, while simultaneously adapting its explanations of these relationships to the user's behavioural data. By training the model with these pairs, the model attempts to capture the contextual cues and patterns that influence app-switching behaviour. To evaluate the effectiveness of different screen text inputs, we train two types of models: one using only the screen text as the input data, and another using both the screen text and the current app name. This model comparison aims to determine whether the additional contextual information from the current app allows the model to make more informed predictions, or if screen text alone provides sufficient context. This fine-tuning approach assesses the value of a targeted input dataset in improving the model's understanding of screen text in relation to next-app prediction.

Additionally, we also train a next-app prediction model for each participant using only their current app as input. We hypothesise that this method will be less effective at predicting the user's next app compared to models that integrate the user's screen text. The prompts for this experiment are provided in Appendix C.

4.1.3. Evaluation method

For each participant, we first assess the performance of the base Llama 3-8b model in predicting the next app a user is likely to open, using this as our baseline. We then evaluate the effectiveness of the model trained only on their current app as input, using this as another baseline measure. For each participant, we filter the dataset to include only their relevant data and use their screen text and current app data as input to the model.

Then, we design an evaluation pipeline to assess the effectiveness of our fine-tuned models in comparison to the base model. For each evaluation run, we specify the participant and the training-evaluation split to ensure that the corresponding fine-tuned model is loaded. We filter the data for only the current participant and further refine it based on the chosen split to include only the days designated for evaluation. For each participant and split, we evaluate three distinct approaches:

- **Screen Text Only:** We include only the screen text as input to the model, which is trained only on screen text data.
- **Screen Text and Current App (Text-Trained Model):** We include both the screen text and the current app as input to the model, which is trained only on screen text data.
- **Screen Text and Current App (Text and App-Trained Model):** We include both the screen text and the current app as input to the model, which is trained on both screen text and current app data.

We subsequently load the fine-tuned model and tokenizer based on the current evaluation method. To evaluate the model's responses, we provide each data entry in the evaluation dataset as a distinct prompt, which is formatted using the fine-tuned tokenizer with no caching between prompts. The predicted next app is extracted from the model's response by removing all attached metadata referring to user roles. The resulting prediction is compared to the actual next app that the participant opened and marked as correct or incorrect. Finally, we obtain an overall prediction accuracy by computing the number of correct predictions out of the total predictions. Each model is initially evaluated once with a temperature of 0 to establish a deterministic accuracy. After this, each evaluation is performed five times using a temperature of 0.6, which follows the default value provided in the base Llama 3 pipeline, adding an element of randomness in response generation to test the model's robustness. The average accuracy is then computed across these five runs.

4.1.4. Model explainability

In addition to evaluating the models for correctness, we randomly sample both correct and incorrect predictions from our fine-tuned models and prompt them for explanations to gain insights into the reasoning behind their responses. By exploring methods to understand the logic of these predictions, we aim to assess the models' ability to provide reasonable explanations of the user's decision-making process, thereby enhancing transparency and interpretability in understanding the models' decisions.

These explanations are generated by providing the model with a screen text entry, the current app, the next app predicted by the model, and asking it to explain its reasoning in a paragraph. This approach helps us validate the model's performance while ensuring that its outputs are understandable and justifiable, even in instances where it is incorrect. Understanding model explanations also resonates with the broader goal of giving end users visibility into how their data drives predictions, which is particularly important when personal textual data is being used. The prompt we use is shown below.

4.2. Experiment 2: ESM activity prediction

Examining the relationship between screen text and real-world activities can enable a more comprehensive understanding of user behaviour on the smartphone and its impact on daily life. Screen text provides insights into the digital content users engage with, which may reflect, influence, or even diverge from their physical activities. Aligning ESM self-reports with screen text allows us to observe how well an LLM can discern real-world user contexts, potentially offering more bespoke interventions or recommendations.

4.2.1. Training method and data preparation

We aim to identify whether a fine-tuned individualised model outperforms a fine-tuned generalised model in predicting real-world activity from smartphone use, or vice versa. To accomplish this, we design a pipeline that trains three distinct models for each participant in our study and evaluates their effectiveness.

We first load the datasets for both the screen text and the ESM responses. Four participants are excluded from the analysis because they completed less than half of the ESM questionnaires. The ESM questionnaire asks participants to report their activities during the five minutes preceding the questionnaire, which are then categorised into one of 10 activity categories following the methodology discussed by Teng et al. (2024a). These activity categories are listed in Appendix B. To align with this, we identify all instances of screen text viewed within the five minutes before they received the questionnaire. As a result, each ESM response is associated with multiple screen text data entries. These can be organised into (screen text) and (ESM response) pairs for input when fine-tuning the model. If a participant did not use their phone in the five minutes preceding their questionnaire response, the entry is discarded. Of the 739 total ESM responses collected across the remaining 17 participants, 189 (25.6%) were discarded under the five-minute phone-use criterion, leaving 550 (74.4%) for analysis. Next, we filter the combined dataset according to the type of model we are training.

Table 2
Conceptual comparison of the three model types for ESM activity prediction.

Model Type	Trained On	Training Data Size	What It Tests
Individualised (80/20)	The user only	Small (own data only)	Can a model learn from just *your* behaviour?
Generalised (LOO)	Everyone else	Large (all other users)	Can a model learn from *lots of other people* instead of you?
Generalised (80/20)	Everyone else	Small (proportional)	Is the advantage due to *better data* or just *more data*?

Table 2 provides a high-level overview of the three model types and the question each is designed to answer. The corresponding training and evaluation splits are defined as follows:

- For the **Individualised (80/20) Model**, we use the first 11 days of the participant's data for training and the last 3 days for evaluation, yielding an approximate 80/20 temporal split (Joseph, 2022).
- For the **Generalised (Leave-One-Out) Model**, we use all data from all other participants for training, and evaluate on the full data from the target participant.
- For the **Generalised (80/20) Model**, we randomly sample training data from other participants such that the training set is four times the size of the participant's evaluation set, preserving the same 80/20 train/test ratio as in the individualised setting, and evaluate on the full data from the target participant.

The resulting (screen text) and (ESM response) pairs are then formatted into prompts for training, as detailed in Section 4.2.2.

4.2.2. Prompt design

To fine-tune the model, we format each data point as a natural language prompt consisting of the screen text followed by a query asking the model to classify the user's real-world activity. This prompt format enables the model to learn mappings between digital content and behaviour categories in an interpretable manner. We apply the same prompt structure across all three model types to enable consistent evaluation. The prompts for this experiment are provided in Appendix C.

4.2.3. Evaluation method

Our evaluation pipeline for assessing the fine-tuned models' effectiveness in predicting user ESM response categories mirrors our approach to next-app prediction evaluation. We begin by evaluating the effectiveness of the base Llama 3-8b model for each participant, which we use as the baseline. Then, for each evaluation run, we specify the participant and model type, ensuring the appropriate fine-tuned model is loaded. We then filter the dataset to include only data relevant to the evaluated participant and method. For each participant, we compare the three model types described above. For each evaluation, we load the corresponding fine-tuned model and tokenizer based on the current method. We then assess the model's performance by formatting each entry in the evaluation dataset as a separate prompt using the fine-tuned tokenizer, with no caching between prompts. The predicted ESM response category is extracted from the model's output by removing all attached metadata referring to user roles. We compare this prediction to the participant's actual ESM response, marking it as correct or incorrect. The overall prediction accuracy is then calculated as the ratio of correct predictions to total predictions. Each model is first evaluated with a temperature of 0 to establish a deterministic accuracy. Following this, evaluations are conducted five times with a temperature of 0.6, and the average accuracy from these five runs is then calculated.

4.2.4. Model explainability

Similar to our next-app prediction approach, we prompt explanations from our fine-tuned model to better understand the reasoning behind the model's responses. We provide the model with a screen text entry and the user's activity category predicted by the model, and ask it to explain its decision in a paragraph. The prompt we use is shown below.

4.3. Experiment 3: in-app activity clustering

While much of smartphone behaviour modelling focuses on app-level interactions, these approaches often overlook the diversity of user activity within individual apps. A single app can support a wide range of behaviours—for instance, one person might use a social media app to chat with friends, while another might use the same app to consume news content or browse professional updates. These intra-app differences are not captured by app name alone, yet they reflect meaningful variation in user intent and experience.

In this experiment, we use screen text to cluster in-app activity patterns, providing a more nuanced view of smartphone use. By summarising and categorising user actions from textual content, we move beyond surface-level logs towards a deeper understanding of behaviour. In doing so, we aim to shift from prediction to explanation by revealing the routines and intentions behind app engagement. Understanding in-app activities also has important implications for applications in health, well-being, and user interface design. For example, distinguishing between active messaging and passive scrolling within the same app could provide more precise indicators of sociability, mood, or attention. Through clustering and activity summarisation, this analysis demonstrates the potential of screen text as a lens into everyday smartphone behaviour, offering interpretive insights that complement the prediction-focused tasks explored in earlier experiments.

4.3.1. Data preparation

Given the large volume and diverse range of screen text that is captured from everyday smartphone use, we aim to summarise these texts to extract what activities users are engaged in. By summarising the text into concise activity descriptions, we can focus on aspects of the text that provide insights into user behaviour while increasing user privacy through data minimisation. We utilise the base Llama 3-8b-Instruct model to summarise these texts. This approach does not require the development or fine-tuning of new models, as the foundation model is already capable of summarising text with high accuracy across various contexts. This pre-trained capability allows for effective generation of concise screen text summaries, enabling us to analyse trends and patterns in the types of information that users engage with.

Our process for summarising the screen text uses a multi-step pipeline. First, we instruct the model to summarise the text from each screen into a few succinct sentences, effectively capturing the core content of what the user has viewed. This summarisation step is essential as it enables us to manage large volumes of text data while retaining key information. Next, we utilise the model again to transform each summary into a list of user activities, breaking down the summary to identify the types of content users are interacting with. For example, one user might be reading a news article on Google Chrome, while another could be browsing online shopping sites. We then ask the model to cluster these activities into groups based on their thematic similarities, aiming to guide the model in finding topics related to user interaction. The topics and their corresponding activities are formatted as a dictionary, where each key represents a topic, and the values are a list of associated words. Two of the authors independently review the full set of generated categories for each of the 10 apps included in the analysis, checking each category against three criteria: semantic coherence (whether the activities grouped under a category are thematically related), non-overlap (whether categories are sufficiently distinct from one another), and behavioural plausibility (whether the category corresponds to a recognisable type of in-app activity rather than an artefact of the underlying text). Disagreements are resolved through discussion, with the model prompted to regenerate categories where both reviewers judge the original grouping to be incoherent or redundant. This approach highlights patterns and relationships between different types of interactions within the same app, providing insights into how various individuals use smartphone apps differently. The prompts used for these summarisation steps are described in Section 4.3.2. Following the categorisation process, each data entry is matched with the topic dictionaries to create a "topics" attribute, which lists the identified topics relevant to the screen text. We also record the duration each screen was viewed for by calculating the time difference between when the user first opened the screen and when they exited it. This multi-step design highlights our emphasis on developing privacy-preserving transformations for textual content, where the summarised screen text conceals raw data and personally identifiable information while retaining sufficient detail to uncover behavioural patterns.

4.3.2. Prompt design

To explore how individuals interact with apps, we designed a series of prompts that utilise the base model's summarisation capabilities. The first prompt focuses on summarising user activity by analysing the screen text and the app it originates from. The second prompt instructs the model to extract significant activities from the summarised user activity. The goal is to identify distinct actions that users perform, prioritising activities involving direct interaction with their devices. The third prompt involves organising the extracted activities into thematic groups. The grouping of these activities under relevant topics can allow for more structured analysis of user interactions and behaviours. The prompts for this experiment are provided in Appendix C.

4.3.3. Analysis methods

Following data preparation and extraction of activities from the screen text, we aggregate the data to calculate the total duration that each participant spent on each activity within individual apps.

We first analyse differences in activity usage across the apps using three metrics:

- **Percentage of Total Time Spent on Each Activity:** For each app, we calculate the percentage of total time users spent on each activity. We normalise usage durations at the participant level to account for variability in total smartphone usage time among participants. This analysis reveals which activities dominate usage within each app.
- **Distribution and Evenness of Activities:** Using Pielou's Evenness Index, we assess how evenly the usage time is distributed across the 10 activities within each app. A higher evenness index indicates that time is spread relatively equally across the activities, while a lower index suggests that a few activities account for most of the app's usage time. This provides insights into whether certain apps are used for a variety of purposes or primarily for specific activities.
- **Participant Engagement in Activities:** For each app and activity, we calculate the number of participants who engaged in the activity. This helps identify how widely each activity is used within the participant pool and whether some activities are more niche compared to others that are universally performed.

We also compute Pearson correlation coefficients across each pair of participants based on their time spent on each activity within each app to assess each individual's degree of similarity in their app usage patterns. This analysis helps uncover whether certain behaviours are consistently shared among users or if there are distinct differences in how individuals engage with these apps. We utilise k-means clustering to group participants based on their time spent on these activities, which can allow for visualisation of similarities across individuals. We restrict the list of apps for our analysis to the 10 most commonly used by participants in the study to ensure meaningful comparisons: Facebook, Facebook

Messenger, Gmail, Google Chrome, Google Maps, Google Messages, Google Photos, Instagram, WhatsApp, and YouTube.

5. Results

5.1. Experiment 1: next-app prediction

The base Llama 3-8b model demonstrates extremely low accuracy, averaging around 0.1% across the 14 days for all participants, indicating that it is not reliable for our domain. Similarly, models trained solely on the current app exhibit low performance, with an average accuracy of approximately 5% over the 14 days for all participants.

Using our models fine-tuned on the screen text, the majority of participants (18 out of 21) achieved the highest performance in the 13/1 split (64% ± 13%) in at least one of the three evaluation approaches, and this was similarly obtained by 16 participants in the 11/3 split (63% ± 12%). All three evaluation approaches showed consistent improvement in accuracy as more training data was included, with the greatest improvement occurring between 1 day (46% ± 16%) and 2 days (58% ± 14%) of training data, as shown in Fig. 1. Prompting the model trained on both the screen text and current app gave the highest overall scores for most participants and performed best across all splits. Overall, using both the screen text and current app as prompts on the model trained only on screen text resulted in the lowest prediction accuracy (56% ± 15%). In contrast, using the same prompts on the model trained with both screen text and app data achieved the highest prediction accuracy (63% ± 15%). However, no significant differences in accuracy were present between the three evaluation approaches, as displayed in Fig. 4.

We also observe several individual characteristics. For example, although most participants show an upward trend in performance as the split ratio increases, some participants, such as P9 and P10, have a peak in performance in the middle split (7/7) followed by a slight decline. Additionally, some participants, such as P4, and P17, show notably higher performance when including both the screen text and current app in their prompts and evaluation model. On the other hand, other participants, such as P5 and P15, perform better when evaluated on models trained only on their screen text.

Given the skewed nature of app usage distributions, where a small number of apps account for the majority of interactions, we additionally compute macro F1 scores to assess performance more equally across all app classes. These results are reported in Appendix E. While macro F1 provides an important fairness-aware perspective, accuracy remains our primary evaluation metric. This is because, in practice, models deployed in real-world settings should prioritise correct predictions on frequently used apps, which often have a greater impact on user experience. Moreover, in-the-wild app usage tends to follow naturally skewed patterns, where certain apps are used far more frequently than others. Therefore, overall accuracy more closely aligns with the system's practical effectiveness.

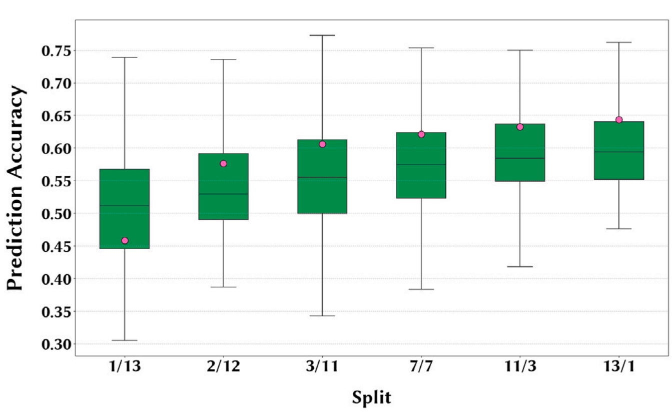

Fig. 1. Next-App Prediction Accuracies by Training/Evaluation Split.

Overall, we observe remarkably consistent results for each participant across both the run with a temperature of 0 and the five runs with a temperature of 0.6 for each evaluation method and split, resulting in a negligible standard deviation. We provide the mean accuracies for each participant, across all evaluation methods and splits, in Appendix D.

Because the next-app task includes cases where the user remains in the current app, we additionally report accuracy separately for same-app continuations and app switches in Figs. 2, 3, 5, 6. The two cases produce noticeably different accuracies. Same-app continuations are predicted with high accuracy across all splits (rising from 69% at the 1/13 split to 84% at the 13/1 split), while app switches are predicted

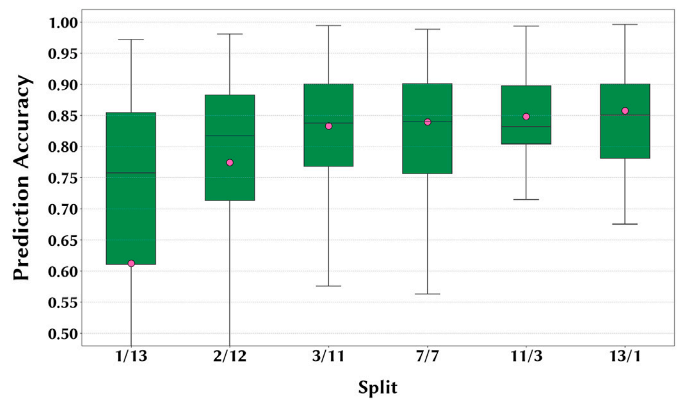

Fig. 2. Next-App Prediction Accuracies by Training/Evaluation Split (Same App).

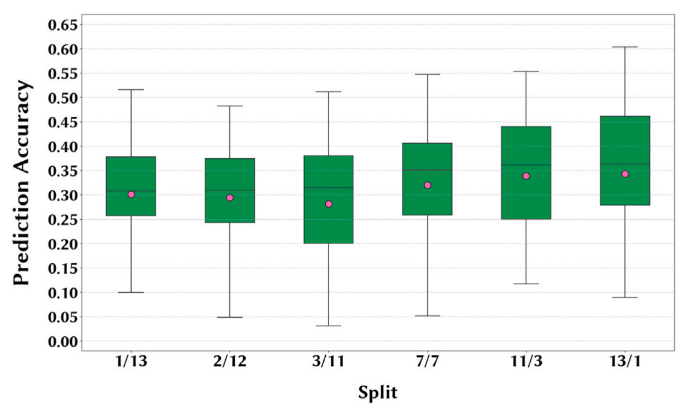

Fig. 3. Next-App Prediction Accuracies by Training/Evaluation Split (App Switch).

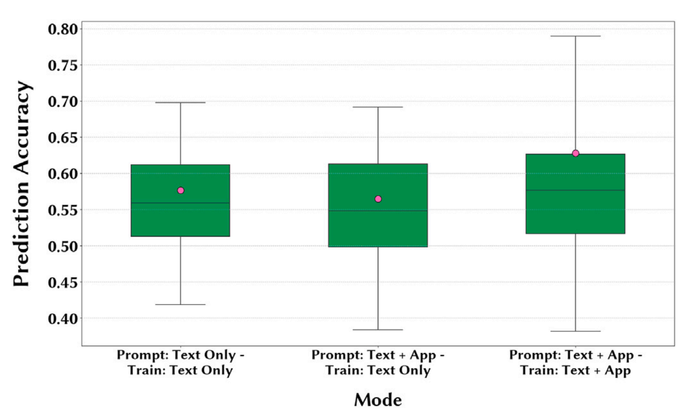

Fig. 4. Next-App Prediction Accuracies by Prompting and Training Mode.

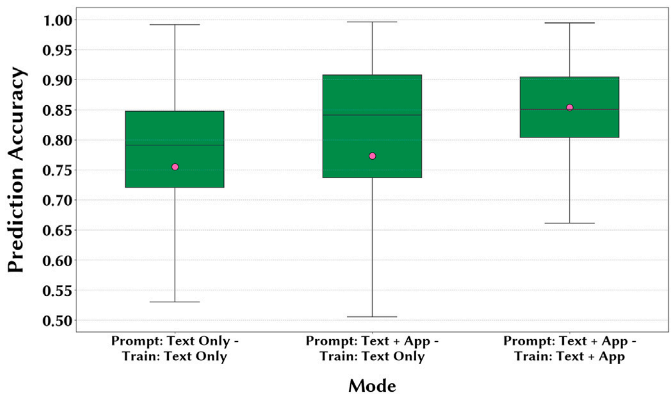

Fig. 5. Next-App Prediction Accuracies by Prompting and Training Mode (Same App).

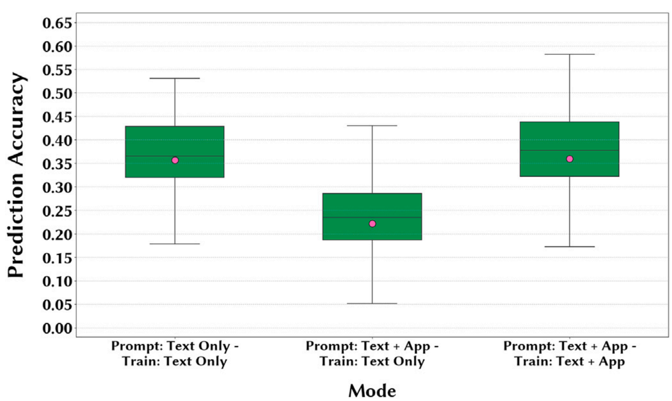

Fig. 6. Next-App Prediction Accuracies by Prompting and Training Mode (App Switch).

Table 3
App prediction model response explanations—correct example.

Screen Text	Current App	Predicted Next App	Actual Next App
Pick Up Now Domino's Carlton (Vic) Add a Voucher Apply THE VOUCHER IS NOT ACCEPTED BY YOUR SELECTED STORE. Domino's Wallet	Domino's App	Google Chrome (✓)	Google Chrome

Explanation: I predicted that the user would go to the **Google Chrome app** next because the **error message on the Domino's app** is likely to **frustrate the user** and prompt them to **seek a solution online**. The user may want to **search for answers or troubleshooting tips** to resolve the issue with the voucher code, and **Google Chrome is a convenient app** to access the internet and search for relevant information. By opening Google Chrome, the user can quickly **search for Domino's customer support, FAQs, or online forums** where they may find a resolution to the issue, making it a likely next app for them to open.

Table 4
App prediction model response explanations – incorrect example.

Screen Text	Current App	Predicted Next App	Actual Next App
How to Politely Stop Your Rambling Co-worker | The Muse Menu	Google Chrome	Slack (✗)	WhatsApp

Explanation: I predicted that the user would next go to the **Slack app** because the text they viewed on **Google Chrome** is an article from **The Muse**, a popular **career development website**, and the topic is **how to politely stop a coworker from rambling**. Given that the user was likely **reading this article at work**, I inferred that they are currently using **Slack, a common workplace communication platform**, to **collaborate with their colleagues**. The user may have read the article, and now wants to **apply the tips to their own work situation**, perhaps by **sending a direct message or making a comment in a channel** to politely interrupt their rambling coworker.

with substantially lower accuracy (29% to 36% across splits). The same pattern holds across prompting conditions, where the best-performing condition for same-app continuations (screen text and current app as prompts on the model trained on both) reaches 84% mean accuracy, while the same condition reaches only 37% on app switches. This indicates that a meaningful portion of the overall accuracy reported above is driven by the model correctly learning when a user will stay in their current app, which is a much more common occurrence in participant behaviour compared to genuine app switches. The app-switch numbers nonetheless remain well above the 5% baseline obtained by the current-app-only model, suggesting that screen text does indicate switching behaviour.

5.1.1. Model explainability

To illustrate how our fine-tuned models determine how the current text a user is viewing may influence their next app choice, we present one example each of a correct and incorrect prediction, along with the model's explanation for its decision, as shown in Tables 3 and 4, respectively.

5.2. Experiment 2: ESM activity prediction

The base Llama 3-8b model produces a very low accuracy, with an average of about 1% across all participants, suggesting that it is not well-suited to our specific data.

We compare three model types for ESM activity prediction: an individualised model trained on a participant's own data (80/20 temporal split), a generalised (leave-one-out) model trained on all data from other participants, and a generalised (80/20) model trained on a volume-matched subset of other-user data to maintain an 80/20 split relative to the target participant's evaluation set size.

The generalised (leave-one-out) models demonstrate a higher overall mean accuracy (87% ± 22%) compared to the individualised models (81% ± 27%). However, this difference is not statistically significant, as indicated by the Wilcoxon signed-rank test, suggesting that both approaches may have comparable overall performance. The results suggest that when trained on large-scale population data, generalised models can perform comparably to personalised models for this task. In contrast, the generalised (80/20) models, which use a limited amount of other-user data matched to the participant's evaluation volume, perform substantially worse (60% ± 23%). A Wilcoxon signed-rank test reveals that the individualised models have significantly higher prediction accuracies compared to the generalised (80/20) models ($W = 14.0$, $p < 0.01$). The median paired difference is 25%, with an interquartile range of 26.3% and a large effect size ($r = 0.76$).

Examining individual participants, we find that the individualised model outperforms the generalised (leave-one-out) model for approximately half of the participants (9 out of 17). Conversely, for the remaining 8 participants, the generalised (leave-one-out) model performs better. Notably, there is an instance of an extreme difference between the two models for one participant. P8 shows a drastic increase in accuracy from 36% with the individualised model to 97% with the generalised (leave-out-one) model and 76% with the generalised (80/20) model. Across all other participants, the generalised (80/20) model performs the worst, and there are no significant differences between the performance of the other two models (Fig. 7).

Similar to our observation with next-app prediction, we find consistent results for each participant across both the run with a temperature

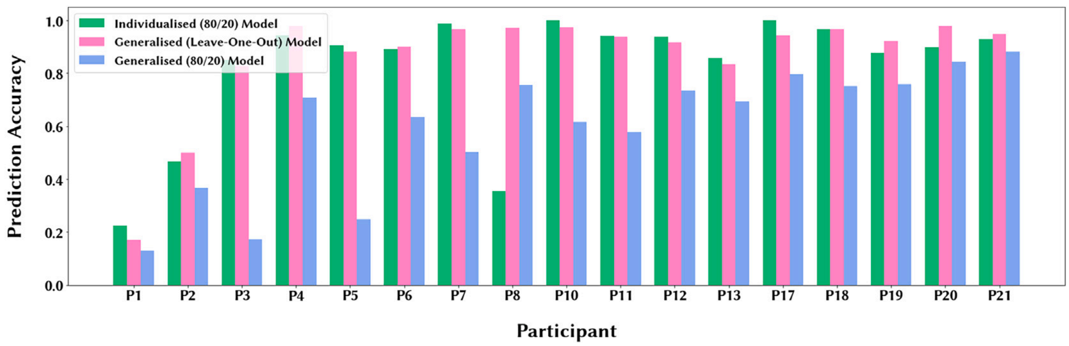

Fig. 7. ESM Activity Prediction Accuracies by Model Type.

of 0 and the five runs with a temperature of 0.6 for each model type, producing a negligible standard deviation.

5.2.1. Model explainability

To demonstrate how our fine-tuned models reason about how the text a user is viewing can predict their current self-reported real-world activity, we provide one example of both a correct and an incorrect prediction, along with the model's rationale for each, as detailed in Tables 5 and 6, respectively.

Table 5
ESM prediction model response explanations—correct example.

Screen Text (Excerpt)	Predicted Activity Category	Actual Activity Category
Indiana Jones and the Dial of Destiny – Rotten Tomatoes navigation bar WHAT TO KNOW Critics Consensus It isn't as thrilling as earlier adventures, but the nostalgic rush of seeing Harrison Ford back in action helps Indiana Jones and the Dial of Destiny find a few final bits of cinematic treasure. Read critic reviews	Entertainment (✓)	Entertainment

Explanation: I predicted the activity category as "**Entertainment**" because the user is currently on the **Rotten Tomatoes website**, which is a popular platform for **reviewing and discovering movies, TV shows, and other forms of entertainment**. The specific page they are on is about a movie, "**Indiana Jones and the Dial of Destiny**", which is a **well-known franchise**. The content on the page includes a **summary, critic reviews, and a navigation bar**, all of which suggest that the user is engaged with the movie's information. Additionally, the presence of a "**Critic Consensus**" section and the mention of "**cinematic treasure**" further reinforce the **entertainment theme**. Overall, the combination of the **website**, the **content**, and the **context** all point to the user being engaged in an **entertainment-related activity**.

Table 6
ESM prediction model response explanations—incorrect example.

Screen Text (Excerpt)	Predicted Activity Category	Actual Activity Category
Time to BeReal. 2 min left to capture a BeReal and see what your friends are up to!	Socialising (✗)	Shopping

Explanation: I predicted the activity category as "**Socialising**" because the text appears to be a notification or prompt from the **BeReal app**, which is a **social media platform** that encourages users to **share unedited, real-life moments**. The language used is **inviting and encouraging**, urging the user to "**capture a BeReal**" and "**see what your friends are up to**", indicating that the app is facilitating **social interaction and community engagement**. The tone is also **casual and informal**, which is consistent with the tone often used in social media platforms. Overall, the text suggests that the user is likely engaged in a **social activity**, specifically using the BeReal app to **share and view content with their friends**.

5.3. Experiment 3: in-app activity exploration

We conduct four analysis tasks to understand usage patterns across different apps: the percentage of time spent on each activity within individual apps, the evenness of activity distributions across apps, the number of participants engaging in each activity, and the relationships between intra-app activities. These analyses provide insights into how users allocate their time across activities, how balanced app usage is across their functionalities, and which activities are most widely adopted by participants.

Analysing the percentage of total time spent on activities within individual apps reveals clear patterns of usage, as shown in Appendix F. Facebook usage is dominated by "Content and Media" which accounts for 68.4% of total time, followed by "Communication" (19.8%), while all other activities contribute less than 7%. Similarly, Instagram usage is concentrated on "Posts and Content" (64.8%) and "User Interaction" (30.0%), reflecting its focus on content creation and engagement. In contrast, apps like Google Maps and Google Chrome exhibit more balanced usage patterns. In Google Maps, time is distributed across "Navigation" (35.4%), "Places and Points of Interest" (25.0%), and "Reviews and Ratings" (15.0%), reflecting its more diverse methods of use. Google Chrome displays a similar distribution, with "Information and Resources" (40.7%) and "General Web Content" (22.7%) being the primary activities, complemented by smaller usage in "Multimedia and Entertainment" (10.6%) and "Shopping and E-commerce" (10.2%). These results highlight the differences between apps that cater to a variety of purposes and those that are more specialised in their usage.

Pielou's Evenness Index provides insights into how evenly time is distributed across activities within each app, illustrated in Fig. 8. Google Maps has the highest evenness score (0.76), indicating that its usage is spread across various features, such as navigation, reviews, and points

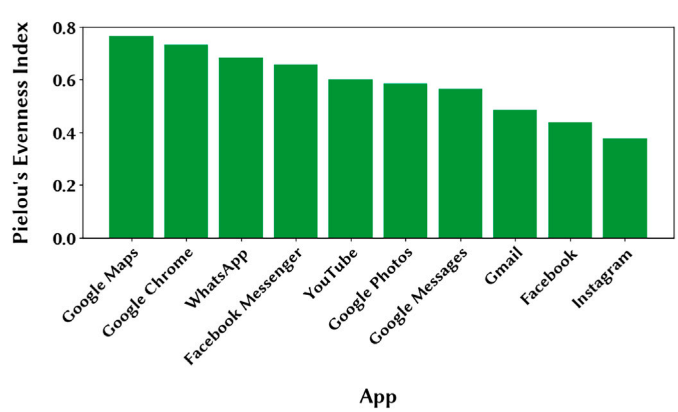

Fig. 8. Activity Usage Distributions.

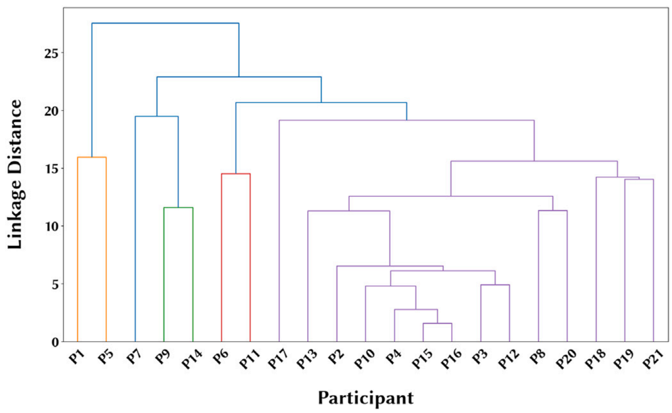

Fig. 9. Participant Hierarchical Clustering.

of interest. Google Chrome (0.73) and WhatsApp (0.68) also display relatively high evenness. Conversely, Instagram has the lowest evenness score (0.38), reflecting its overwhelming focus on "Posts and Content" and "User Interaction", with limited use of other features. Facebook (0.44) and Gmail (0.48) also show low evenness, demonstrating a concentration on specific activities like messaging and email. These findings align with the results of the previous task.

The number of participants engaging in each activity reveals additional information about usage breadth, as provided in Appendix G. Gmail and Google Chrome see high engagement across key activities, with 20 participants using "Emails and Messages" in Gmail and 19 participants accessing "General Web Content" and "Information and Resources" in Chrome. Google Maps also demonstrates widespread engagement, with 19 participants exploring "Places and Points of Interest" and performing "Search and Query" tasks. In contrast, Facebook and Instagram show narrower engagement, with fewer participants accessing peripheral activities. For instance, only 4 participants engage with "Job and Career" on Facebook, while Facebook Messenger's "Marketplace and Shopping" activity is accessed by just 8 participants. Similar to the previous tasks, these patterns suggest that while some apps have features widely adopted by most participants, others are dominated by core functionalities with limited use of less central activities.

We also compare activity usage duration among participants using Pearson's correlation coefficient. By calculating these correlation coefficients, we can identify which participants exhibit similar behaviours in app usage. These correlations can then be used for hierarchical clustering, allowing us to group participants based on the degree of similarity in their behaviours using k-means clustering with Euclidean distance and visualise these relationships through a dendrogram, as shown in Fig. 9. These visualisations allow for clearer identification of participant groups and can highlight individuals who may exhibit unique characteristics or behaviours. We cluster our participants into 6 clusters based on their app usage characteristics, with their behavioural patterns described in Table 7.

To extend our analysis beyond individual apps, we examine correlations between in-app activity patterns across different apps. We do this by calculating the time spent on every app-activity pair for each participant. We then compute the Pearson correlation for each app-activity pair across all participants. Table 8 lists examples of strong observed correlations, where increased engagement in one activity is consistently associated with increased engagement in the other.

Table 7
Participant Clusters and Activity Usage Characteristics.

Cluster	Participants	Description
Cluster 1 – Everyday Enthusiasts	P2, P3, P4, P8, P10, P12, P13, P15, P16, P18, P19, P20, P21	Typical usage patterns across a range of apps
Cluster 2 – Video Voyagers	P9, P14	Consumes large amounts of video content across multiple platforms (Facebook, Google Chrome, Instagram, YouTube), low usage of communication-focused apps (WhatsApp)
Cluster 3 – Feed Fanatics	P6, P11	Uses social media heavily to view posts (Facebook, Instagram) and chat with others (Facebook Messenger, Instagram, WhatsApp), does not use productivity apps much (Gmail)
Cluster 4 – Surfing Shopper	P7	Spends a lot of time online shopping (Google Chrome) and viewing related ads (YouTube)
Cluster 5 – Browsing Brainiac	P17	Primarily uses only communication apps (Facebook Messenger, Google Messages, WhatsApp) to chat with others, makes a lot of web searches and views news content (Google Chrome), does not view social media posts at all
Cluster 6 – Social Streamers	P1, P5	Prevalent social media usage for messaging others and viewing social media posts (Instagram, WhatsApp), spends long durations watching videos and reading comments (YouTube)

Table 8
Examples of high Inter-App activity correlations across Participants.

App 1 Activity	App 2 Activity	Correlation
Gmail: News and Updates	WhatsApp: Calls and Audio	0.986
Gmail: Invitations and Events	WhatsApp: Calls and Audio	0.970
Google Maps: Navigation	WhatsApp: Reactions and Emojis	0.921
Gmail: Transactions and Financials	WhatsApp: Media and Attachments	0.902
Facebook: Financial and Transactions	Instagram: Promotions and Advertisements	0.886
Facebook: Surveys and Feedback	Instagram: Promotions and Advertisements	0.876
Gmail: Invitations and Events	WhatsApp: Content and Posts	0.862
Instagram: Events and Locations	YouTube: Community and Social	0.856
Facebook Messenger: Marketplace and Shopping	Google Chrome: Shopping and E-Commerce	0.839
Google Chrome: Shopping and E-Commerce	WhatsApp: Links and Websites	0.815

6. Discussion

6.1. Experiment 1: next-app prediction

We aim to predict the next app a user is likely to open based on their smartphone screen text and current app usage. Our findings indicate that screen text can be effective for next-app prediction, and incorporating both screen text and current app information into the model further improves prediction accuracy. This reinforces the idea that personal context derived from the specific text a user engages with can enhance a model's ability to anticipate user behaviour (Peters et al., 2024). Beyond accuracy, our approach emphasises why a particular prediction might be made, providing richer insight into the underlying factors that drive a user's decision-making process. We note, however, that the aggregate accuracy figures conflate two qualitatively different tasks: predicting that a user will remain in their current app, and predicting a true app switch.

The results show that the majority of participants achieve their highest prediction accuracy with the 13/1 split, followed closely by the 11/3 split. This follows the notion that a greater amount of training data generally leads to better model performance in predictive modelling (Junqué de Fortuny et al., 2013). The sharp improvement observed between the one-day and two-day training splits highlights the importance of providing the model with sufficient data to capture temporal patterns in user behaviour. This finding suggests that a one-day snapshot of user interactions is likely insufficient for uncovering deeper preferences, whereas even a modest extension to two days can yield substantially improved performance in personalised settings (Munappy et al., 2022). Our comparison of different prompting methods reveals that models trained on both screen text and current app data outperform those trained only on screen text. This suggests that the additional contextual cues provided by the current app can be beneficial in enhancing next-app predictions (Zhao et al., 2019b). However, the lack of significant differences across the three evaluation approaches indicates that screen text alone can be highly informative, especially with sufficient data volume.

The individual differences we observe among participants further illustrate the impact of fine-tuning. While most participants show improved performance with increased training data, some exhibit peak performance with a smaller amount of training, suggesting that too much data might cause overfitting (Montesinos López et al., 2022). Additionally, the variability in performance based on whether the current app is included in the prompts suggests that its predictive value is dependent on how much additional context it provides. In some cases, the screen text alone may already indicate the current app, making its inclusion redundant or even introducing noise if the model is not trained on this information. For some participants, the current app provides important context that significantly enhances prediction accuracy, while for others, the screen text alone is a sufficient indicator of their next used app, suggesting that optimal strategies may vary across users.

Another important aspect lies in the LLM's ability to provide explanations for their next-app predictions. As illustrated in Tables 3 and 4, the fine-tuned model attempts to rationalise its decision based on the contextual cues present in the screen text and the user's current app. Importantly, these explanations offer interpretive insights that can help users, developers, and researchers understand the reasoning behind a prediction. By doing so, we move beyond purely end result-focused modelling towards a framework where user behaviour becomes more transparent and comprehensible. Even though the model can still make incorrect predictions, such as predicting Slack when the user actually switched to WhatsApp, the rationale behind its decision highlights the thought process that led to the error (Eigner and Händler, 2024). This transparency could facilitate more informed refinements to the model, such as ensuring it leverages a more representative dataset, or incorporates additional contextual features from the data to better capture the nuances of each user's behaviour. Moreover, displaying these self-explanations to users can potentially enhance trust, as individuals are more likely to adopt predictive technologies when they can see and understand the reasoning process behind these suggestions (Wanner et al., 2022).

Overall, our results indicate the feasibility of leveraging screen text for next-app prediction and highlight the importance of customisable models that adapt to individual usage patterns. By incorporating on-device training or hybrid approaches that protect screen text privacy, models could dynamically tune the amount of contextual information used based on observed behaviour, further refining next-app prediction while remaining applicable across diverse user groups. Alongside these predictions, the explanations derived from these predictive models serve as a valuable resource for promoting trust, as users gain a clearer sense of why specific apps are recommended. Developers and researchers similarly benefit by identifying opportunities to optimise recommendation models, reduce biases, and design more intuitive interfaces, aiding the creation of more personalised and transparent recommendation systems.

6.2. Experiment 2: ESM activity prediction

Although predicting real-world activities from screen text is a complex task, we observe that there is strong potential for gaining deeper insight into user behaviour through their smartphone usage. Across the three model types – individualised, generalised (leave-one-out), and generalised (80/20) – we find that both individualised and generalised (leave-one-out) models achieve high overall accuracy, with no statistically significant difference between them. This suggests that when trained on large-scale data from other users, generalised models can approximate or even match the performance of personalised models for many individuals. This is promising for real-world deployment scenarios where personalisation may not be feasible, and large, aggregated training datasets are available.

However, when we control for training data volume by using the generalised (80/20) model, where the training set is matched to a four-to-one ratio (80/20 split) relative to each participant's test set in a way similar to the individualised condition, we observe a substantial and statistically significant performance gap in favour of the individualised models. On average, the individualised models outperform the filtered generalised models by over 20 percentage points in accuracy, with a large effect size. This indicates that when data is constrained, training on a participant's own behavioural patterns is more effective than using an equivalent amount of data from other users.

Notably, the generalised model significantly outperforms the individualised model for P8, suggesting that a limited or unrepresentative personal dataset may limit the effectiveness of personalisation (Campagner et al., 2023). Conversely, the poor performance of all models for P1 suggests that certain activities might be inherently more challenging to predict based on screen text alone. This could be due to a variety of factors, such as the ambiguity of the text, less distinct patterns of smartphone use, or real-world activities that do not strongly correlate with digital behaviour (Baumel and Yom-Tov, 2018). However, the consistency in results across all the runs for each participant and model type, with negligible standard deviation, suggests overall robustness of our models. This stability is essential for developing reliable and reproducible models, and we further recommend conducting multiple iterations when evaluating models to ensure their robustness (Raschka, 2018). These results point to a trade-off between scalability and accuracy. Generalised models trained on large datasets offer practical benefits for scalable deployment, especially in contexts where collecting personal data is not viable. However, in domains where personalisation is critical, such as mental health monitoring, education, or digital interventions, investing in individualisation can provide meaningful gains in predictive performance (Johnson et al., 2020).

In addition to numeric accuracy, our models also support interpretive insights through generated explanations. Taken together with the quantitative results, this highlights the utility of screen text for bridging digital interactions and real-world experiences.

Our study also highlights how LLMs can provide interpretive insights into real-world activity predictions by explaining their decisions. Unlike the next-app prediction task, where explanations predominantly focused on in-app cues and immediate user intent, the rationales here focus on bridging a user's digital interactions with their daily offline behaviours. These explanations can uncover subtle linkages between activities, which can be especially informative in contexts where numerical accuracy does not fully capture a user's lived experience. These include behaviours such as a series of calendar notifications before reporting "Working", or reading film reviews preceding "Entertainment". Tables 5 and 6 show how the fine-tuned model interprets screen text to infer activity categories, highlighting how seemingly routine smartphone interactions can shed light on underlying real-world behaviour (Nakamura, 2015). These patterns reinforce that screen text contains signals relevant to users' self-reported contexts, and that LLMs are capable of surfacing such associations through both classification and narrative explanation. This interpretive ability can be particularly valuable in domains like health and well-being, where understanding the digital cues preceding certain activities could inform early intervention or tailored support for individuals (Olawade et al., 2024). By examining these rationales, designers and researchers can not only refine prediction algorithms but also gain insight into the personal nuances of user behaviours that standard accuracy metrics might obscure. Additionally, this emphasis on transparency allows predictions to be assessed, making it easier to detect biases and incorrect assumptions in the model while promoting user trust in data-driven interventions.

Overall, these findings suggest that both individualised and generalised models have value, but in different contexts. Therefore, it may be worthwhile to explore hybrid approaches where model training is conducted on a mix of data specific to the user and more general, readily available data collected over time from other users. Additionally, investigating the specific characteristics of users who benefit more from personalisation could contribute towards exploring the balance between individualised and generalised predictive models.

6.3. Experiment 3: in-app activity exploration

Our findings demonstrate the applicability of screen text in understanding in-app activity usage. This level of behavioural granularity offers a much deeper and more nuanced view of user behaviour than simply tracking app usage, which often fails to capture the diverse ways in which an app can be used (Brinberg et al., 2021). For instance, as our study suggests, two individuals might both use a social media app like Instagram, but one might be primarily engaged in messaging friends, while the other is more focused on reading posts. Screen text analysis allows us to distinguish these different modes of interaction, and this information can enhance existing research across various use cases.

One of the primary benefits of being able to analyse in-app activities through screen text is the potential for more personalised and context-aware interactions. For example, identifying high levels of passive social media consumption might indicate low mood (Zubair et al., 2023), while frequent messaging activity could signal strong social support networks (Yue et al., 2023). Such distinctions allow healthcare professionals or well-being applications to offer more targeted advice or interventions (Tong et al., 2021). Additionally, this capability could enhance the development of smarter, more intuitive digital assistants and bespoke features within apps.

Our study shows that the ability to categorise in-app activities and cluster individuals based on screen text can also contribute to a more detailed understanding of user habits. This insight not only informs app design but can also enhance user experience when using smartphones. Developers can leverage this data to identify underutilised areas of their app or features that fall short of expected engagement with certain demographics. By analysing how often these patterns occur within each app session, they can design new features that align more closely with user interaction patterns (da Silva et al., 2022). Additionally, clustering users based on their behaviour allows for identifying groups of similar users and enables relevant customisation of recommendation algorithms (Yalcin and Bilge, 2021). By recognising patterns among users with similar habits, recommendations can be made that align more closely with their preferences, while also offering recommendations of new content or features that may appeal to users with different usage patterns (Kulkarni and Rodd, 2020). This approach can allow for more precise and personalised recommendations, leading to an enhanced user experience.

In addition to within-app usage patterns, our analysis reveals strong behavioural correlations across different apps, suggesting that certain activity types tend to be engaged with at similar levels. For example, users who frequently engage with Gmail activities such as "News and Updates" or "Invitations and Events" also tend to show high engagement with WhatsApp activities like "Calls and Audio" and "Media and Attachments." These associations may reflect shared behavioural tendencies, suggesting that users often engage with complementary informational and social content across platforms. Similarly, the strong correlation between "Google Maps: Navigation" and "WhatsApp: Reactions and Emojis" may point to a shared context of mobility and communication, such as reacting to location updates or coordinating plans. We also observe consistent cross-platform links between commerce-related activities, such as "Facebook: Financial and Transactions" and "Instagram: Promotions and Advertisements," or "Facebook Messenger: Marketplace and Shopping" and "Google Chrome: Shopping and E-Commerce." These examples suggest that users may pursue similar goals across different digital platforms, and that screen text-based analysis can reveal these multi-app routines.

An additional consideration is the role of data minimisation and summarisation in reducing privacy exposure without undermining model performance. Rather than transmitting or storing raw screen text, developers and researchers could implement summarisation or partial anonymisation techniques to obfuscate sensitive details while retaining key behavioural indicators. However, summarisation may not guarantee privacy preservation. Due to poor instruction-following or hallucinations, LLM-generated summaries could still retain names, inferred routines, or even information about third parties who did not consent to the study. Summarisation should therefore be treated as one component of a broader privacy-preserving pipeline, most usefully when combined with on-device processing and explicit redaction of sensitive entities, rather than as a privacy guarantee in itself. Balancing fidelity and protection is a critical challenge in smartphone sensing studies with large volumes of data (Harari et al., 2016). If summarised data can approximate the semantic meaning of the raw text, it could assist in developing more user-trusted data pipelines, particularly in highly personal domains such as health and well-being. As summarisation capabilities and on-device LLMs mature, researchers may increasingly opt for partial data exposures, aligning with growing ethical and regulatory demands for data minimisation (Patil et al., 2025).

Our results demonstrate the potential to move beyond surface-level app usage data to understand the specifics of user behaviour, leading to more tailored experiences across different domains. By finding the right balance between the volume of information captured and privacy safeguards, it becomes possible to harness the richness of screen text for advanced behavioural analysis while ensuring that users maintain confidence in how their personal data is handled.

6.4. Ethical, privacy, and methodological considerations

6.4.1. Privacy and ethics

Screen text is inherently sensitive, as it can easily contain personal or confidential information. Therefore, the use of such data requires a careful approach to ensure that user privacy is preserved at all stages of data collection, analysis, and storage.

Our findings suggest that for certain behaviour prediction tasks, a generalised model trained on aggregated data may perform comparably to an individualised model for a large portion of the population. Relying on a generalised approach can greatly reduce the need for individuals to share extensive personal screen text data, thereby limiting potential privacy risks. However, for some users with more distinctive or less typical patterns of smartphone use, such as P7 and P17, an individualised model could yield meaningful gains in prediction accuracy. In those cases, our results show that collecting just two days of an individual's data substantially improves model performance compared to using only one day's data, rising from 46% to 58%. This improvement is beneficial because it demonstrates meaningful gains in accuracy without requiring extensive data collection over prolonged periods.

One mitigation strategy for the privacy concerns associated with screen text data is the utilisation of on-device LLMs and analysis tools (Xu et al., 2023). These tools perform all computations directly on the user's device, ensuring that sensitive data does not leave the user's local environment. By keeping data processing and analysis entirely on the device, the risk of data breaches, unauthorised access, or misuse by third parties is significantly reduced (Peng et al., 2024). This approach allows users to retain full control over their data, which is particularly important when dealing with personal and potentially sensitive information such as screen text. The trend towards on-device processing aligns with broader developments in privacy-preserving machine learning, which seek to empower users by decentralising data processing (Abramson et al., 2020). This method minimises the need to transmit data to remote servers, thereby mitigating vulnerabilities associated with data transfer and storage in cloud environments. Additionally, on-device LLMs can deliver real-time interventions and personalised recommendations at reduced latency and without compromising user privacy. This functionality is crucial in applications requiring timeliness, such as in health monitoring (Li et al., 2024b).

However, user autonomy and consent remain vital. Users should be given clear instructions to opt-in to on-device processing and be provided with options to manage or delete their data based on their preferences (Kreuter et al., 2018). Additionally, the potential for errors exists in predicting user behaviour using their screen text. Misinterpretations or inaccuracies in the data analysis could lead to incorrect assumptions about a user's intentions, behaviours, or preferences. Such errors could have negative consequences, especially in sensitive applications like well-being or health interventions. For instance, an incorrect prediction might result in misleading advice, which could negatively impact a user's decision-making (P.S. D.V., 2023). Therefore, it is essential to ensure that users are fully informed about the nature of the data being collected, how it will be used, and the limitations of the predictive models.

Finally, we caution against treating LLM-based summarisation as a guarantee for privacy preservation. Although summarisation reduces the volume of text retained and can omit verbatim content, the resulting summaries may still encode sensitive information. The privacy benefit of summarisation therefore depends on what the summarisation step is explicitly instructed to omit, where it is performed (on-device versus in the cloud), and what downstream systems do with the output.

6.4.2. LLM scalability

While LLMs offer significant potential for enhancing predictive models, their scalability remains a challenge, given the resource-intensive nature of fine-tuning (Patil and Gudivada, 2024). As participant numbers grow, so do computational demands, creating challenges for large-scale research on LLM fine-tuning. Furthermore, although on-device LLMs offer a platform for privacy-preserving analysis, the limited processing power of smartphones constrains their capabilities compared to more powerful, cloud-based models (Li et al., 2024a). Although smaller on-device models, like Gemma 2B (Google, 2024), showcase advancements, they may not yet reach the performance levels of models presented in our study. This limitation raises questions about the practicality of relying solely on on-device LLMs, particularly in scenarios where high accuracy and low latency are critical.

A promising avenue is adopting hybrid approaches that split tasks between on-device and cloud-based computation (Ding et al., 2024). Using this approach, initial data processing could be performed locally on the user's device to ensure that sensitive data remains private. This initial processing could handle tasks that require immediate, real-time analysis or involve data that should not be transmitted off-device due to privacy concerns. For more complex or resource-intensive analyses, the system could then selectively offload specific tasks to the cloud, using data that has been preprocessed and filtered for personal information. This would allow the utilisation of the greater computational power and advanced capabilities of cloud-based models without compromising the privacy of the most sensitive data (Li et al., 2024a). This flexibility could enable the system to optimise both privacy and performance, providing a framework for further developments in LLM scalability and privacy-preserving technologies.

6.4.3. Evaluating LLM explanations

Evaluating the reliability and trustworthiness of explanations generated by LLMs poses a significant challenge in the current landscape of explainable AI (Zhou et al., 2024). While a correct prediction might suggest that the explanation is also accurate, this correlation can potentially be misleading (Nauta et al., 2023). A model could arrive at a correct prediction based on cues not specifically related to the rationale it provides, thereby undermining the credibility of the explanation itself (Lyu et al., 2022). Moreover, explanations can sometimes be overly generic or influenced by training data biases, causing them to appear plausible without necessarily revealing the true decision path taken by the model.

One approach to mitigating these issues is to involve users in manual reviews of model outputs and reasonings, such as asking them to state whether the model's explanations of their actions match their actual behaviours (Hassija et al., 2023). Although human-in-the-loop evaluations can be valuable for spotting inconsistencies or logical fallacies, they are inherently resource-intensive and difficult to scale. In addition, recall bias may lead to inaccurate recollection of a user's actions, which could affect the validity of these evaluations (Althubaiti, 2016). This variability highlights the need for more robust, standardised metrics and benchmarking protocols that can assess the consistency and truthfulness of explanations, while emphasising the importance of caution in their use.

6.4.4. Limits of LLM-generated rationales

A consideration throughout this work is that LLM-generated rationales are model outputs, not verified evidence of user intention. It is useful to distinguish between three things that such a rationale could in principle do: explain the model's prediction by surfacing the cues it attended to, generate a plausible post-hoc interpretation of the behaviour visible in the screen text, or recover the user's actual intention. Our experiments give us reasonable confidence in the first two, but not necessarily in the third. For example, the Domino's to Google Chrome rationale in Table 3 is a plausible reading of why a user might switch apps after a voucher error, but we did not ask the participant whether frustration was in fact the trigger. The incorrect-prediction example in Table 4 makes the point in reverse, where the model's rationale for predicting Slack is internally consistent (workplace article to workplace messaging app), but the user actually switched to WhatsApp. The coherence of a rationale is therefore not necessarily evidence of its correctness, and a fluent explanation may accompany an incorrect prediction.

Therefore, LLM-generated rationales should be used as hypothesis generation, rather than as strict, ground-truth accounts of user behaviour. Validating them would require user studies in which participants are asked to confirm or correct the model's interpretation of their own actions. Although human-in-the-loop evaluations can spot inconsistencies, they are resource-intensive and subject to recall bias, highlighting the need for standardised faithfulness benchmarks before LLM-generated interpretations can support consequential decisions in domains like health, education, or well-being.

6.5. Broader implications

Across the three experiments presented in this study, we demonstrate the value of screen text for modelling smartphone-based behaviour. These findings collectively suggest several broader insights. First, screen text provides a nuanced, semantically rich representation of user context that supports prediction of both digital actions and real-world activities. Second, beyond predictive performance, LLM-generated explanations offer a window into the underlying behavioural patterns reflected in screen interactions, helping to make such models more interpretable and allowing for an enhanced understanding of how people use their smartphones. Third, the comparison between generalised and individualised models reveals a trade-off between scalability and personalisation: while generalised models may suffice in realistic deployment scenarios, individualised models are especially effective when user behaviour is unique or training data is constrained. Lastly, the ability to detect fine-grained in-app activities through screen text further expands the potential for personalised, context-aware systems, with applications ranging from well-being interventions to more responsive app design.

Together, these findings highlight the potential of screen text not just as a predictive feature, but as a window into users' behaviours, routines, and intentions. By combining predictive modelling with natural language explanations, we take a step towards more interpretable, context-aware systems that move beyond surface-level app usage to offer deeper insights into how people engage with their smartphones. In doing so, this work contributes to the broader goal of shifting from prediction to explanation in digital behaviour modelling and understanding smartphone use.

7. Limitations and future work

The screen text sensor used for collecting the data in our study is still relatively new, and its full capabilities have not yet been fully explored. As our research is exploratory in nature and we are in the early stages of designing and testing methods for analysing this type of data, there is room for improvement in our methodology. Future work could involve experimenting with different models, refining data preprocessing techniques, and exploring alternative approaches to data collection, potentially leading to more insightful findings across different domains. Our two-week collection window also limits what we can conclude about how user-specific language patterns evolve over time. Individual vocabularies, app repertoires, and behavioural routines can shift over months in ways that a short field study cannot capture, and longitudinal screen text data would allow future work to examine how fine-tuned models drift and whether continual learning approaches can keep individualised models aligned with a user's evolving behaviour.

Our current approach largely treats screen text as a uniform input stream and does not yet differentiate between distinct types of content, such as user-inserted text versus received or consumed content. Future work could involve extracting such structured features from screen text to distinguish between content types and interaction modes. This could be further enhanced by fusing screen text with other sensors, such as keyboard entry or physical activity, to better contextualise user behaviour. For example, combining long-duration passive reading behaviour with location or time-of-day data could improve understanding of routines or attention states. We view this study as a first step towards such integrative modelling, focusing on assessing the potential of raw screen text semantics and providing a foundation for incorporating more complex feature engineering.

Another limitation of our work is that we do not conduct a systematic analysis of the computational or energy overhead from continuously capturing screen text and processing it with an LLM. While our focus was on the feasibility of screen text-based LLM-driven behaviour analysis rather than resource consumption, future research could examine device battery usage and processing demands. Existing measurements from smartphone sensors, such as accelerometers and gyroscopes, have been extensively documented (Anagnostopoulos et al., 2017). Our preliminary analysis reveals that the data volume of screen text is approximately 25% of what is typically recorded by accelerometers or gyroscopes over the same time period. Nevertheless, additional empirical evaluations are needed to help researchers and developers balance overhead with the benefits of improved behavioural explainability.

Our implementation uses server-side processing for screen text analysis, which introduces ethical concerns due to the transmission of potentially sensitive data. While this approach is necessary for supporting experimentation and fine-tuning in a research setting, it is not ideal for real-world deployment. Future work should aim at adopting on-device processing techniques that enable all analysis to be performed locally. Recent advances in lightweight LLMs and privacy-preserving architectures make this shift increasingly feasible, and would allow users to retain control over their data while reducing risks associated with remote storage and transmission.

Although all three of our experiments could be extended in both data collection and analysis, the ESM response prediction experiment, in particular, offers strong potential for future work. In the dataset we used, the ESM responses are limited to what participants were doing in the last five minutes. However, future studies could involve designing more comprehensive ESM questionnaires that target various aspects of participants' daily lives. For example, questions could be used to assess participants' current mental state, productivity levels, or emotional well-being (Chan et al., 2018). Incorporating these additional dimensions would provide a richer understanding of self-reported user behaviour and allow us to explore how screen text can be used to predict a broader range of psychological and behavioural states. This could significantly enhance the applicability of our methods in domains such as mental health, productivity enhancement, and personalised digital interventions.

A critical challenge in explainable AI is establishing reliable and trustworthy LLM-based explanations, particularly when seemingly convincing rationales may not reflect a model's actual decision path. Developing robust evaluation frameworks, such as gold-standard datasets containing large samples of user actions on their devices, could help measure explanation fidelity within smartphone and user behaviour studies. Future research could also incorporate user studies to explore how system-generated explanations of user activities influence their trust and comprehension (Falconnet et al., 2023). By integrating interpretability-focused methods and real-world feedback, we can move closer to developing models for understanding human behaviour that are both accurate and transparent in their decision-making.

8. Conclusion

This paper presents novel methods for analysing screen text data captured from smartphones to enhance the understanding of user behaviour. Through our experiments, we illustrate how leveraging screen text to fine-tune LLMs can offer nuanced insights into in-app behaviour. Our findings also show that the integration of screen text and LLM fine-tuning goes beyond conventional behavioural analysis methods by highlighting the reasons behind user actions, which is important across several domains such as health and well-being and education. However, effectively balancing rich data collection with privacy and scalability remains a challenge. Techniques such as data minimisation and summarisation show promise for safeguarding sensitive information without sacrificing the depth of behavioural insights. Future work should explore hybrid approaches that incorporate both user-specific and general datasets, refine explanations for model predictions, and extend screen text analysis to broader populations and contexts. As smartphone use continues to grow, the ability to effectively analyse screen text will play an increasingly important role in deepening our understanding of user behaviour.

CRediT authorship contribution statement

Songyan Teng: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Hong Jia:** Writing – review & editing, Methodology, Conceptualization. **Simon D'Alfonso:** Writing – review & editing, Methodology, Conceptualization. **Vassilis Kostakos:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

Appendix A. Screen text de-duplication algorithm

Our de-duplication algorithm processes a sequence of screen text entries to identify and group continuous segments of text viewed by a user.

The algorithm starts with an empty string and compares each new screen text entry to the current string. Specifically, it checks if the end of the current string overlaps with the beginning of the next screen text entry. If an overlap is found, the algorithm appends only the new, previously unseen text to the current string. A new text string is started either when no overlap is detected or if the app in use changes. Each text string, along with its viewing duration and the combined de-duplicated text, is then recorded. An example of how the de-duplication algorithm is applied to a user's viewed text as they scroll is shown in Table A.1.

Table A.1
Screen Text De-Duplication Example.

Entry No.	Original Screen Text	Matches End?	De-Duplicated Screen Text
1	Welcome to our website.	✗	Welcome to our
2	to our website. Please sign in.	✓	website. Please sign in.
3	Please sign in or register.	✗	Please sign in or register.
4	Latest news updates.	✗	Latest news updates.
5	news updates. Click here for more.	✓	Click here for more.
6	updates. Click here for more.	✓	
7	Breaking: Major event.	✗	Breaking: Major event.
8	event. Details to follow.	✓	Details to follow.
9	Details to follow in the next article.	✗	Details to follow in the next article.
10	Exclusive interview inside.	✗	Exclusive interview
11	interview inside. Read now.	✓	inside. Read now.

Appendix B. ESM activity categories

See Table B.1

Table B.1
ESM activity categories and example activities.

Category	Examples
Eating/Drinking	Having Lunch, Drinking Tea, Eating Out
Entertainment	Watching TV, Gaming, Reading
Exercising	Walking, Gym, Rock Climbing
Housework	Cooking, Cleaning, Laundry
Online Searching	Searching Real Estate, Viewing Online Maps
Resting	Relaxing, In Bed, Napping
Shopping	Getting Groceries, Online Shopping
Socialising	Calling, Visiting Friends, Attending Convention
Travelling	Catching the Train, Driving, Biking
Working	Work Meeting, Studying For Test, Conducting Experiments

Appendix C. Model prompts

Experiment 1: next-app prediction

Screen Text Only
Instruction: Here is some text that a user has viewed on their smartphone. What app are they most likely to go to next? Answer with only the name of the app.

Text: [one screen of text viewed by the user]

Next App: [the user's next used app]

Screen Text + Current App
Instruction: Here is some text that a user has viewed on their smartphone and the current app they are using. What app are they most likely to go to next? Answer with only the name of the app.

Text: [one screen of text viewed by the user]

Current App: [the app where this text was generated]

Next App: [the user's next used app]

Current App Only
Instruction: Here is the current app that a user is using on their smartphone. What app are they most likely to go to next? Answer with only the name of the app.

Current App: [the user's current app]

Next App: [the user's next used app]

Next-App Prediction Explanation
Instruction: Here is some text that a user has viewed on their smartphone and the current app they are using. This is the next app you predicted they would go to.

Text: [one screen of text viewed by the user]

Current App: [the app where this text was generated]

Next App: [the model's prediction of the user's next used app]

Please explain why you made this prediction in one paragraph.

Experiment 2: ESM activity prediction

ESM Activity Prediction
Instruction: Here is some text that a user has viewed on their smartphone. What category does their current activity most likely belong to? Answer with only the name of the activity category.

Text: [one screen of text viewed by the user]

Activity Category: [the user's activity category]

ESM Activity Prediction Explanation
Instruction: Here is some text that a user has viewed on their smartphone. This is the category of the current activity you predicted for the user.

Text: [one screen of text viewed by the user]

Activity Category: [the model's prediction of the user's current activity category]

Please explain why you made this prediction in one paragraph.

Experiment 3: in-app activity exploration

Screen Text Summarisation
Instruction: Here is some text that a user has viewed on their smartphone and the app they are on. Summarise what the user is doing and how they are interacting with their phone based on the text and the app the text is from. Respond with only this summary and nothing else.

Text: [one screen of text viewed by the user]

App: [the app where this text was generated]

Summary: [summary of the text]

Activity Extraction
Instruction: Analyse the following summary of a user's smartphone activity, including the text they've viewed and the app they were using. Extract up to three distinct, significant activities…

Text: [summary of text viewed by the user]

App: [the app where this text was generated]

Activities: [extracted activities from the text summary]

Activity Grouping
Instruction: Here is a list of words related to activities on a smartphone app. Group each word under a topic related to user interaction and format the result as a dictionary…

Words: [activities extracted from app use]

App: [the app where these activities occurred]

Topics: [dictionary containing the topics and grouped words]

Appendix D. Next-app prediction accuracies

See Table D.1

Table D.1
Next-App prediction Accuracies by Mode and Split.

Participant	Mode	Split: 1/13	Split: 2/12	Split: 3/11	Split: 7/7	Split: 11/3	Split: 13/1
P1	Prompt: Text Only – Train: Text Only	0.4753	0.5268	0.5726	0.5913	0.6153	**0.6270**
	Prompt: Text + App – Train: Text Only	0.4122	0.5175	0.5182	0.5168	**0.5202**	0.4945
	Prompt: Text + App – Train: Text + App	0.5539	0.5723	0.5997	0.6174	**0.6353**	0.6290
P2	Prompt: Text Only – Train: Text Only	0.1826	0.9046	0.9075	0.9044	0.9211	**0.9470**
	Prompt: Text + App – Train: Text Only	0.1614	0.8906	0.8974	0.8933	0.9027	**0.9316**
	Prompt: Text + App – Train: Text + App	0.1533	0.8894	0.9141	0.9070	0.9238	**0.9501**
P3	Prompt: Text Only – Train: Text Only	0.4455	0.4479	0.4541	0.4449	0.5198	**0.5484**
	Prompt: Text + App – Train: Text Only	0.4607	0.4978	0.5316	0.5406	0.5635	**0.5646**
	Prompt: Text + App – Train: Text + App	0.5670	0.4591	0.5389	0.5852	**0.6013**	0.5224
P4	Prompt: Text Only – Train: Text Only	0.4716	0.5287	0.5401	0.5702	0.5838	**0.6039**
	Prompt: Text + App – Train: Text Only	0.5922	0.5939	0.6228	0.6216	**0.6913**	0.6716
	Prompt: Text + App – Train: Text + App	0.6372	0.6237	0.6110	0.6199	0.6262	**0.6987**
P5	Prompt: Text Only – Train: Text Only	0.5147	0.5010	0.4921	0.5541	0.5559	**0.5636**
	Prompt: Text + App – Train: Text Only	0.5106	0.4950	0.5142	0.5526	0.5624	**0.5761**
	Prompt: Text + App – Train: Text + App	0.4210	0.4093	0.4786	0.4665	0.5161	**0.5395**
P6	Prompt: Text Only – Train: Text Only	0.4485	**0.5031**	0.4615	0.4911	0.4184	0.4760
	Prompt: Text + App – Train: Text Only	0.5261	0.5458	0.5573	**0.5912**	0.5551	0.5360
	Prompt: Text + App – Train: Text + App	0.5179	0.4005	**0.5760**	0.5549	0.5380	0.5661
P7	Prompt: Text Only – Train: Text Only	0.5092	0.5025	0.4350	0.5000	**0.5145**	0.5072
	Prompt: Text + App – Train: Text Only	0.2008	0.2628	0.3159	0.3835	0.5227	**0.5476**
	Prompt: Text + App – Train: Text + App	0.3335	0.4325	0.4620	0.5238	**0.5342**	0.4947
P8	Prompt: Text Only – Train: Text Only	0.5118	0.5293	0.5324	0.5531	**0.5599**	0.5550
	Prompt: Text + App – Train: Text Only	0.4459	0.4856	0.5145	0.5112	0.5488	**0.5701**
	Prompt: Text + App – Train: Text + App	0.5653	0.5515	0.5681	**0.5938**	0.5772	0.5850
P9	Prompt: Text Only – Train: Text Only	0.5731	0.6100	0.6215	**0.6257**	0.6080	0.6036
	Prompt: Text + App – Train: Text Only	0.3053	0.5189	0.6682	**0.6838**	0.6428	0.6632
	Prompt: Text + App – Train: Text + App	0.7387	0.7358	0.7447	0.7536	0.7497	**0.7620**
P10	Prompt: Text Only – Train: Text Only	0.4790	0.5067	0.5333	0.5222	**0.5676**	0.5607
	Prompt: Text + App – Train: Text Only	0.5844	0.5941	0.5848	**0.6002**	0.5958	0.5950
	Prompt: Text + App – Train: Text + App	0.6356	0.6297	0.6587	**0.6667**	0.6507	0.6377
P11	Prompt: Text Only – Train: Text Only	0.6121	0.6267	0.6443	**0.6464**	0.6377	0.5939
	Prompt: Text + App – Train: Text Only	0.5447	0.5622	0.6014	0.6620	**0.6697**	0.6634
	Prompt: Text + App – Train: Text + App	0.4539	0.4549	0.4506	0.4557	0.5657	**0.5890**
P12	Prompt: Text Only – Train: Text Only	**0.6140**	0.5886	0.6030	0.5577	0.5991	0.6025
	Prompt: Text + App – Train: Text Only	0.5138	0.4975	0.4746	0.4177	**0.5217**	0.5072
	Prompt: Text + App – Train: Text + App	0.3817	0.3869	0.3874	0.5030	0.5795	**0.6143**

(*continued on next page*)

Table D.1 *(continued)*

Participant	Mode	Split: 1/13	Split: 2/12	Split: 3/11	Split: 7/7	Split: 11/3	Split: 13/1
P13	Prompt: Text Only – Train: Text Only	0.4912	0.5122	0.5508	0.5604	0.5814	**0.6316**
	Prompt: Text + App – Train: Text Only	0.4160	0.4783	0.5360	0.5379	0.5043	**0.5679**
	Prompt: Text + App – Train: Text + App	0.5644	0.5831	0.5957	0.6189	0.5991	**0.6430**
P14	Prompt: Text Only – Train: Text Only	0.4932	**0.5749**	0.5253	0.5448	0.5515	0.5188
	Prompt: Text + App – Train: Text Only	0.4892	0.5285	0.5086	0.4987	**0.5410**	0.4957
	Prompt: Text + App – Train: Text + App	0.4445	0.4641	0.4706	0.4355	0.4606	**0.5159**
P15	Prompt: Text Only – Train: Text Only	0.6246	0.6267	0.6335	0.6342	**0.6624**	0.6368
	Prompt: Text + App – Train: Text Only	0.6588	0.6748	0.6613	**0.6793**	0.6705	0.6608
	Prompt: Text + App – Train: Text + App	0.5642	**0.5665**	0.5028	0.5633	0.4928	0.5145
P16	Prompt: Text Only – Train: Text Only	0.5673	0.6060	0.6443	0.6766	**0.6844**	0.6459
	Prompt: Text + App – Train: Text Only	0.4568	0.4635	0.4566	0.4658	0.5799	**0.5802**
	Prompt: Text + App – Train: Text + App	0.5230	0.5546	0.5525	0.5692	**0.6077**	0.5827
P17	Prompt: Text Only – Train: Text Only	0.4256	0.4801	0.5337	0.5386	**0.5493**	0.5151
	Prompt: Text + App – Train: Text Only	0.2165	0.4491	**0.6815**	0.6711	0.5461	0.5256
	Prompt: Text + App – Train: Text + App	0.7244	0.7539	0.7728	0.7893	0.7729	**0.7998**
P18	Prompt: Text Only – Train: Text Only	0.3342	0.3358	0.3427	0.4480	0.5472	**0.5814**
	Prompt: Text + App – Train: Text Only	0.5065	0.5499	0.5591	0.5885	**0.6178**	0.6001
	Prompt: Text + App – Train: Text + App	0.5089	0.4980	0.4959	0.6046	0.5957	**0.6071**
P19	Prompt: Text Only – Train: Text Only	0.5716	0.5889	0.6029	0.6246	0.6533	**0.6930**
	Prompt: Text + App – Train: Text Only	0.4329	0.4704	0.4876	0.4574	0.5836	**0.6876**
	Prompt: Text + App – Train: Text + App	0.5527	0.5807	0.5715	0.5924	**0.6027**	0.5721
P20	Prompt: Text Only – Train: Text Only	0.5350	0.5738	0.5966	0.6067	**0.6211**	0.6095
	Prompt: Text + App – Train: Text Only	0.5183	0.4959	0.4961	0.6152	0.6067	**0.6361**
	Prompt: Text + App – Train: Text + App	0.6100	0.6267	0.6258	0.6457	0.6741	**0.7047**
P21	Prompt: Text Only – Train: Text Only	0.5399	0.5471	0.5550	0.5300	0.6167	**0.6973**
	Prompt: Text + App – Train: Text Only	0.5846	0.6058	0.6148	0.6232	**0.6412**	0.6356
	Prompt: Text + App – Train: Text + App	0.5690	0.5780	**0.5836**	0.5748	0.5743	0.5830

Appendix E. Next-app prediction macro F1 scores

See Table E.1

Table E.1
Next-App prediction macro F1 scores by Mode and Split.

Participant	Mode	Split: 1/13	Split: 2/12	Split: 3/11	Split: 7/7	Split: 11/3	Split: 13/1
P1	Prompt: Text Only – Train: Text Only	0.0296	0.0845	0.1671	0.1281	0.3148	**0.4021**
	Prompt: Text + App – Train: Text Only	0.0337	0.1422	0.2249	0.2836	0.2839	**0.3664**
	Prompt: Text + App – Train: Text + App	0.2138	0.2988	0.3380	0.3484	**0.4498**	0.4326
P2	Prompt: Text Only – Train: Text Only	0.1035	0.1182	0.1310	0.2939	0.3826	**0.4568**
	Prompt: Text + App – Train: Text Only	0.0705	0.1603	0.3348	0.3215	0.3524	**0.4503**
	Prompt: Text + App – Train: Text + App	0.2958	0.3274	0.3692	0.3892	0.5035	**0.5865**
P3	Prompt: Text Only – Train: Text Only	0.1199	0.1405	0.1585	0.2018	**0.2596**	0.2443
	Prompt: Text + App – Train: Text Only	0.1477	0.1820	0.2487	0.1816	0.2452	**0.3156**
	Prompt: Text + App – Train: Text + App	0.3134	0.3418	0.3329	0.3804	0.3578	**0.4038**
P4	Prompt: Text Only – Train: Text Only	0.1128	0.1949	0.1903	0.1972	0.3366	**0.3786**
	Prompt: Text + App – Train: Text Only	0.1425	0.2402	0.2912	0.3060	0.3625	**0.4212**
	Prompt: Text + App – Train: Text + App	0.3186	0.3181	0.3071	0.3748	0.4183	**0.4630**
P5	Prompt: Text Only – Train: Text Only	0.0808	0.1243	0.1435	0.1801	0.2443	**0.3305**
	Prompt: Text + App – Train: Text Only	0.1436	0.1739	0.2356	0.1747	0.1441	**0.2594**
	Prompt: Text + App – Train: Text + App	0.3159	0.3900	0.3465	0.3702	**0.4440**	0.4171
P6	Prompt: Text Only – Train: Text Only	0.1096	0.1072	0.1626	0.2338	0.3376	**0.3614**
	Prompt: Text + App – Train: Text Only	0.1746	0.2197	0.2203	0.2676	0.1739	**0.2965**
	Prompt: Text + App – Train: Text + App	0.3519	0.3876	0.3813	0.3930	0.4320	**0.4470**
P7	Prompt: Text Only – Train: Text Only	0.1443	0.1850	0.1683	0.2137	0.2858	**0.3000**
	Prompt: Text + App – Train: Text Only	0.1714	**0.2736**	0.2450	0.2508	0.2600	0.2695
	Prompt: Text + App – Train: Text + App	0.3016	0.3100	0.2816	0.3423	0.4157	**0.4195**
P8	Prompt: Text Only – Train: Text Only	0.1898	0.1879	0.2066	**0.3229**	0.3004	0.3150
	Prompt: Text + App – Train: Text Only	0.1701	0.1958	0.3084	**0.4064**	0.3487	0.3113
	Prompt: Text + App – Train: Text + App	0.3671	0.3678	0.3743	0.4504	0.4603	**0.4882**
P9	Prompt: Text Only – Train: Text Only	0.0791	0.1357	0.1609	0.2491	0.2404	**0.3606**
	Prompt: Text + App – Train: Text Only	0.0321	0.0813	0.2834	0.3199	0.2555	**0.3679**
	Prompt: Text + App – Train: Text + App	0.4330	0.3995	0.4300	**0.4661**	0.4090	0.4612

(*continued on next page*)

Table E.1 *(continued)*

Participant	Mode	Split: 1/13	Split: 2/12	Split: 3/11	Split: 7/7	Split: 11/3	Split: 13/1
P10	Prompt: Text Only – Train: Text Only	0.1505	0.1131	0.2664	0.2689	**0.3499**	0.3283
	Prompt: Text + App – Train: Text Only	0.2184	0.1777	0.3686	0.2568	**0.4199**	0.2986
	Prompt: Text + App – Train: Text + App	0.3027	0.2829	0.3549	0.4167	0.4663	**0.4868**
P11	Prompt: Text Only – Train: Text Only	0.1838	0.1851	0.2620	0.3188	0.3885	**0.5329**
	Prompt: Text + App – Train: Text Only	0.2435	0.2724	0.2883	0.3358	0.2075	**0.4345**
	Prompt: Text + App – Train: Text + App	0.3623	0.3696	0.3695	0.4139	0.5013	**0.5831**
P12	Prompt: Text Only – Train: Text Only	0.1244	0.1970	0.2367	0.3217	0.3338	**0.4121**
	Prompt: Text + App – Train: Text Only	0.1881	0.1507	0.2533	**0.3238**	0.3193	0.3080
	Prompt: Text + App – Train: Text + App	0.3778	0.3847	0.4078	0.4575	0.4878	**0.4967**
P13	Prompt: Text Only – Train: Text Only	0.0594	0.0876	0.1317	0.2224	0.2811	**0.3408**
	Prompt: Text + App – Train: Text Only	0.0405	0.0902	0.2817	0.3045	0.2330	**0.3304**
	Prompt: Text + App – Train: Text + App	0.1937	0.3081	0.3597	0.4008	**0.4076**	0.3994
P14	Prompt: Text Only – Train: Text Only	0.0585	0.1286	0.2418	0.2327	0.3316	**0.3416**
	Prompt: Text + App – Train: Text Only	0.1016	0.1350	0.2490	0.2115	0.0536	**0.3129**
	Prompt: Text + App – Train: Text + App	0.1413	0.3295	0.3147	0.3555	**0.4664**	0.4397
P15	Prompt: Text Only – Train: Text Only	0.1562	0.1724	0.1392	0.2086	**0.2661**	0.2106
	Prompt: Text + App – Train: Text Only	0.1670	0.1949	0.1788	0.1429	**0.2974**	0.2117
	Prompt: Text + App – Train: Text + App	0.2488	0.2597	0.2171	**0.3025**	0.2391	0.2413
P16	Prompt: Text Only – Train: Text Only	0.1638	0.1308	0.1494	0.1390	0.2105	**0.2995**
	Prompt: Text + App – Train: Text Only	0.1145	0.1124	0.1658	0.1526	0.1878	**0.1897**
	Prompt: Text + App – Train: Text + App	0.2827	0.3590	0.3719	**0.3859**	0.3845	0.3292
P17	Prompt: Text Only – Train: Text Only	0.1385	0.2199	0.2963	0.3500	0.4024	**0.4319**
	Prompt: Text + App – Train: Text Only	0.0766	0.1478	0.3842	0.4821	0.4133	**0.5008**
	Prompt: Text + App – Train: Text + App	0.3575	0.4116	0.4638	0.5345	0.5320	**0.6305**
P18	Prompt: Text Only – Train: Text Only	0.0528	0.0733	0.1168	0.1256	0.2757	**0.3266**
	Prompt: Text + App – Train: Text Only	0.0224	0.0450	0.0912	0.0841	0.2952	**0.3583**
	Prompt: Text + App – Train: Text + App	0.1928	0.2389	0.2321	0.3005	0.3860	**0.4036**
P19	Prompt: Text Only – Train: Text Only	0.0600	0.0896	0.1071	0.2802	0.3169	**0.3390**
	Prompt: Text + App – Train: Text Only	0.0454	0.1330	0.2065	**0.3433**	0.3245	0.3258
	Prompt: Text + App – Train: Text + App	0.2802	0.3212	0.3330	0.4252	**0.4595**	0.4044
P20	Prompt: Text Only – Train: Text Only	0.1260	0.1550	0.1541	0.1666	0.3065	**0.4191**
	Prompt: Text + App – Train: Text Only	0.1542	0.1746	0.1608	0.1700	0.2249	**0.3870**
	Prompt: Text + App – Train: Text + App	0.3628	0.3750	0.3466	0.3812	0.4021	**0.4385**
P21	Prompt: Text Only – Train: Text Only	0.1186	0.2228	0.2625	0.2760	0.4068	**0.4581**
	Prompt: Text + App – Train: Text Only	0.1745	0.2927	0.2711	0.3033	0.3725	**0.4071**
	Prompt: Text + App – Train: Text + App	0.2813	0.2980	0.3198	0.3719	**0.4828**	0.4736

Appendix F. Percentage of time spent on each activity in each app

See Fig. F.1.

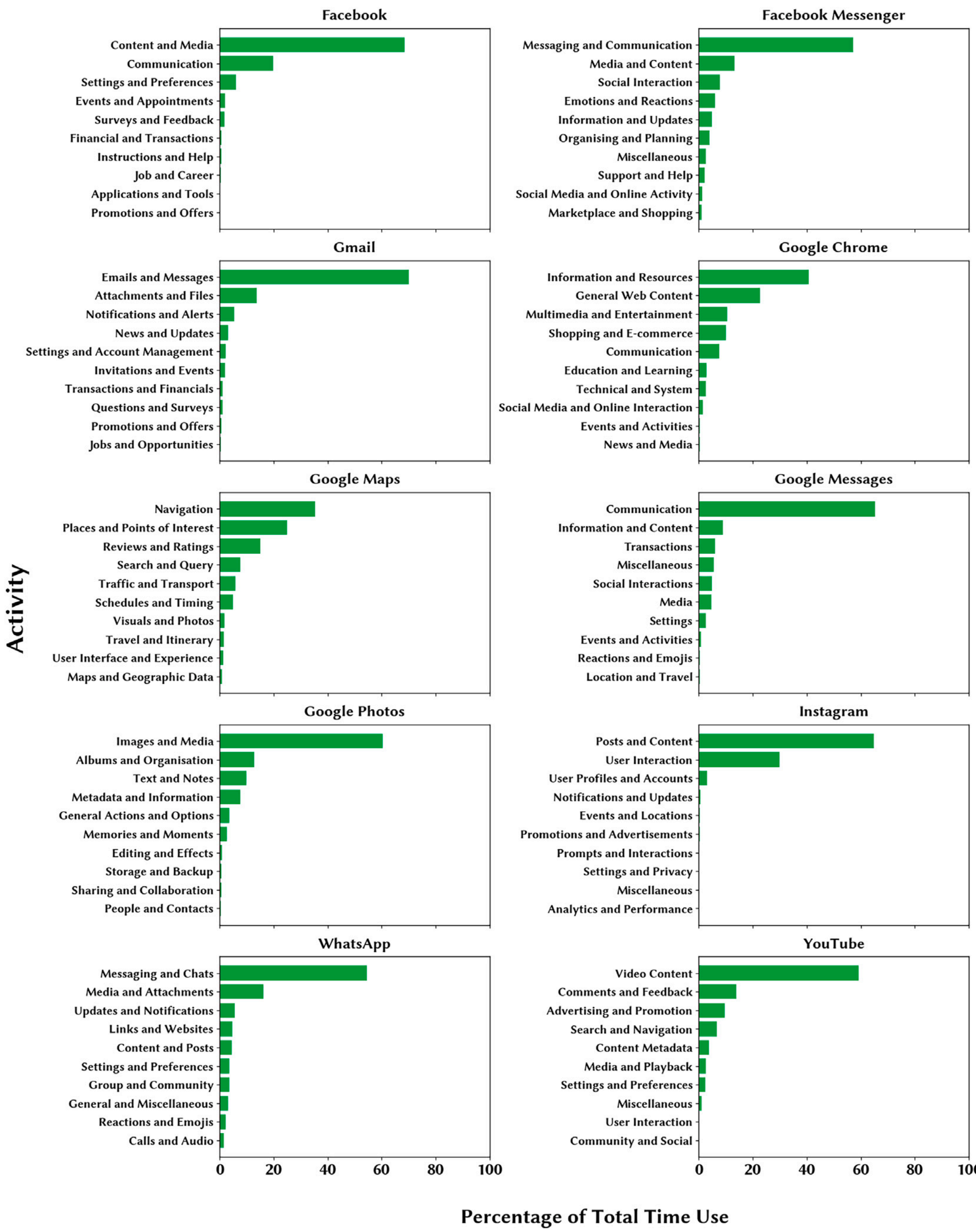

Fig. F.1. Percentage of time spent on each activity.

Appendix G. Number of participants who used each activity in each app

See Fig. G.1.

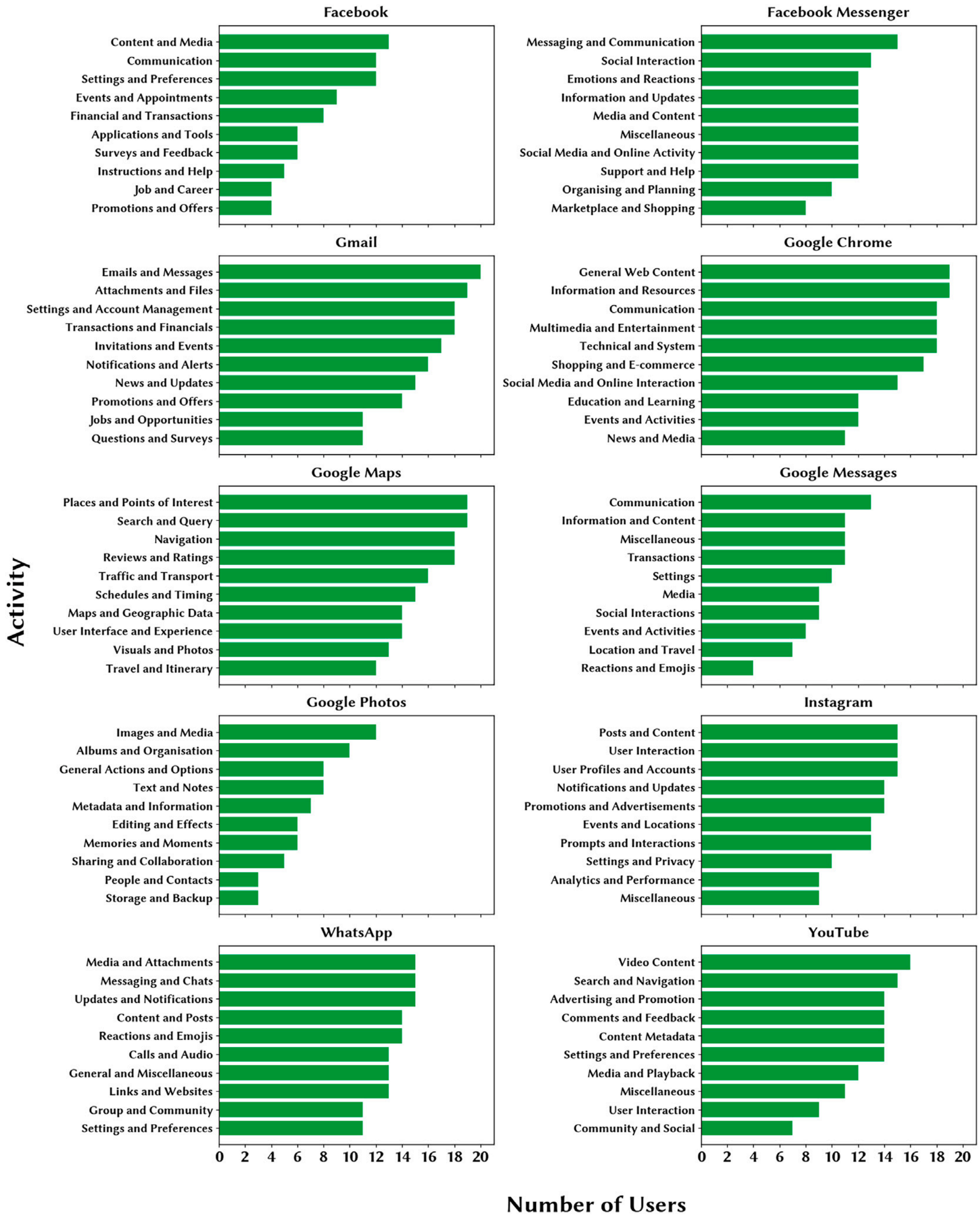

Fig. G.1. Number of participants who used each activity.

Data availability

The data that has been used is confidential.

References

Abramson, W., Hall, A.J., Papadopoulos, P., Pitropakis, N., Buchanan, W.J., 2020. A Distributed Trust Framework for Privacy-Preserving Machine Learning. Springer International Publishing, pp. 205–220. https://doi.org/10.1007/978-3-030-58986-8

Althubaiti, A., 2016. Information bias in health research: definition, pitfalls, and adjustment methods. J. Multidiscip. Healthc. 211, https://doi.org/10.2147/jmdh.s104807

Anagnostopoulos, T., Garcia, J.C., Goncalves, J., Ferreira, D., Hosio, S., Kostakos, V., 2017. Environmental exposure assessment using indoor/outdoor detection on smartphones. Pers. Ubiquitous Comput. 21, 761–773. https://doi.org/10.1007/s00779-017-1028-y

Baumel, A., Yom-Tov, E., 2018. Predicting user adherence to behavioral ehealth interventions in the real world: examining which aspects of intervention design matter most. Transl. Behav. Med. 8, 793–798. https://doi.org/10.1093/tbm/ibx037

Beames, J.R., Han, J., Shvetcov, A., Zheng, W.Y., Slade, A., Dabash, O., Rosenberg, J., O'Dea, B., Kasturi, S., Hoon, L., Whitton, A.E., Christensen, H., Newby, J.M., 2024. Use of smartphone sensor data in detecting and predicting depression and anxiety in young people (12–25 years): a scoping review. Heliyon 10, e35472. https://doi.org/10.1016/j.heliyon.2024.e35472

Beiwinkel, T., Kindermann, S., Maier, A., Kerl, C., Moock, J., Barbian, G., Rössler, W., 2016. Using smartphones to monitor bipolar disorder symptoms: a pilot study. JMIR Ment. Health 3, e2. https://doi.org/10.2196/mental.4560
Böhmer, M., Hecht, B., Schöning, J., Krüger, A., Bauer, G., 2011. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. ACM, https://doi.org/10.1145/2037373.2037383
Brinberg, M., Ram, N., Yang, X., Cho, M.J., Sundar, S.S., Robinson, T.N., Reeves, B., 2021. The idiosyncracies of everyday digital lives: using the human screenome project to study user behavior on smartphones. Comput. Hum. Behav. 114, 106570. https://doi.org/10.1016/j.chb.2020.106570
Brown, N.B., 2024. Enhancing trust in LLMs: algorithms for comparing and interpreting LLMs. arxiv:2406.01943.
Campagner, A., Famiglini, L., Carobene, A., Cabitza, F., 2023. Everything is varied: the surprising impact of instantial variation on ML reliability. Appl. Soft Comput. 146, 110644. https://doi.org/10.1016/j.asoc.2023.110644
Cao, H., Lin, M., 2017. Mining smartphone data for app usage prediction and recommendations: a survey. Pervasive Mob. Comput. 37, 1–22. https://doi.org/10.1016/j.pmcj.2017.01.007
Chan, L., Swain, V.D., Kelley, C., de Barbaro, K., Abowd, G.D., Wilcox, L., 2018. Students' experiences with ecological momentary assessment tools to report on emotional well-being. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2. pp. 1–20, https://doi.org/10.1145/3191735
Cheung, M., 2024. A reality check of the benefits of LLM in business. arxiv:2406.10249.
Ciman, M., Wac, K., Gaggi, O., 2015. iSenseStress: assessing stress through human-smartphone interaction analysis. In: Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare. ICST, https://doi.org/10.4108/icst.pervasivehealth.2015.259280
Cornet, V.P., Holden, R.J., 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. J. Biomed. Inform. 77, 120–132. https://doi.org/10.1016/j.jbi.2017.12.008
da Silva, L.F., Parreira Junior, P.A., Freire, A.P., 2022. Mobile user interaction design patterns: a systematic mapping study. Information 13, 236. https://doi.org/10.3390/info13050236
de Vries, L.P., Baselmans, B.M.L., Bartels, M., 2020. Smartphone-based ecological momentary assessment of well-being: a systematic review and recommendations for future studies. J. Happiness Stud. 22, 2361–2408. https://doi.org/10.1007/s10902-020-00324-7
Ding, Y., Niu, C., Wu, F., Tang, S., Lyu, C., Chen, G., 2024. Enhancing on-device LLM inference with historical cloud-based LLM interactions. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, ACM, pp. 597–608, https://doi.org/10.1145/3637528.3671679
Eigner, E., Händler, T., 2024. Determinants of LLM-assisted decision-making. arxiv:2402.17385.
Falconnet, A., Coursaris, C.K., Beringer, J., Van Osch, W., Sénécal, S., Léger, P.M., 2023. Improving user experience with recommender systems by informing the design of recommendation messages. Appl. Sci. 13, 2706. https://doi.org/10.3390/app13042706
Fang, L., Zhang, S., Jia, H., Goncalves, J., Kostakos, V., 2024. Screentk: seamless detection of time-killing moments using continuous mobile screen text and on-device LLM. arxiv:2407.03063.
Ferrara, E., 2024. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: a survey of early trends, datasets, and challenges. Sensors 24, 5045. https://doi.org/10.3390/s24155045
Ferreira, D., Goncalves, J., Kostakos, V., Barkhuus, L., Dey, A.K., 2014. Contextual experience sampling of mobile application micro-usage. In: Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services. ACM, https://doi.org/10.1145/2628363.2628367
Ferreira, D., Kostakos, V., Dey, A.K., 2015. AWARE: mobile context instrumentation framework. Front. ICT 2, https://doi.org/10.3389/fict.2015.00006
Fulford, D., Mote, J., Gonzalez, R., Abplanalp, S., Zhang, Y., Luckenbaugh, J., Onnela, J.P., Busso, C., Gard, D.E., 2021. Smartphone sensing of social interactions in people with and without schizophrenia. J. Psychiatr. Res. 137, 613–620. https://doi.org/10.1016/j.jpsychires.2020.11.002
Goodyear, V.A., Armour, K.M., Wood, H., 2018. Young people learning about health: the role of apps and wearable devices. Learn. Media Technol. 44, 193–210. https://doi.org/10.1080/17439884.2019.1539011
Google, 2024. Gemma: open models based on gemini research and technology. arxiv:2403.08295.
Goslen, A., Kim, Y.J., Rowe, J., Lester, J., 2024. LLM-based student plan generation for adaptive scaffolding in game-based learning environments. Int. J. Artif. Intell. Educ.. https://doi.org/10.1007/s40593-024-00421-1
Guo, B., Ouyang, Y., Guo, T., Cao, L., Yu, Z., 2019. Enhancing mobile app user understanding and marketing with heterogeneous crowdsourced data: a review. IEEE Access 7, 68557–68571. https://doi.org/10.1109/access.2019.2918325
Han, D., Han, M., 2024. Unsloth AI. https://github.com/unslothai/unsloth.
Harari, G.M., Lane, N.D., Wang, R., Crosier, B.S., Campbell, A.T., Gosling, S.D., 2016. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect. Psychol. Sci. 11, 838–854. https://doi.org/10.1177/1745691616650285
Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A., 2023. Interpreting black-box models: a review on explainable artificial intelligence. Cogn. Comput. 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8
Huang, Y., Xiong, H., Leach, K., Zhang, Y., Chow, P., Fua, K., Teachman, B.A., Barnes, L.E., 2016. Assessing social anxiety using GPS trajectories and point-of-interest data. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, https://doi.org/10.1145/2971648.2971761
Jager, K.J., Tripepi, G., Chesnaye, N.C., Dekker, F.W., Zoccali, C., Stel, V.S., 2020. Where to look for the most frequent biases? Nephrology 25, 435–441. https://doi.org/10.1111/nep.13706
Johnson, K.B., Wei, W., Weeraratne, D., Frisse, M.E., Misulis, K., Rhee, K., Zhao, J., Snowdon, J.L., 2020. Precision medicine, AI, and the future of personalized health care. Clin. Transl. Sci. 14, 86–93. https://doi.org/10.1111/cts.12884
Joseph, V.R., 2022. Optimal ratio for data splitting. Stat. Anal. Data Min.: ASA Data Sci. J. 15, 531–538. https://doi.org/10.1002/sam.11583
Junqué de Fortuny, E., Martens, D., Provost, F., 2013. Predictive modeling with big data: is bigger really better? Big Data 1, 215–226. https://doi.org/10.1089/big.2013.0037
Karim, F., Oyewande, A., Abdalla, L.F., Chaudhry Ehsanullah, R., Khan, S., 2020. Social media use and its connection to mental health: a systematic review. Cureus. https://doi.org/10.7759/cureus.8627
Katsarou, K., Yu, G., Beierle, F., 2022. Whatsnextapp: LSTM-based next-app prediction with app usage sequences. IEEE Access 10, 18233–18247. https://doi.org/10.1109/access.2022.3150874
Ke, L., Tong, S., Cheng, P., Peng, K., 2024. Exploring the frontiers of LLMs in psychological applications: a comprehensive review. arxiv:2401.01519.
Khaokaew, Y., Xue, H., Salim, F.D., 2024. Maple: mobile app prediction leveraging large language model embeddings. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8. pp. 1–25, https://doi.org/10.1145/3643514
Kreuter, F., Haas, G.C., Keusch, F., Bähr, S., Trappmann, M., 2018. Collecting survey and smartphone sensor data with an app: opportunities and challenges around privacy and informed consent. Soc. Sci. Comput. Rev. 38, 533–549. https://doi.org/10.1177/0894439318816389
Kucirkova, N., Gerard, L., Linn, M.C., 2021. Designing personalised instruction: a research and design framework. Br. J. Educ. Technol. 52, 1839–1861. https://doi.org/10.1111/bjet.13119
Kulkarni, S., Rodd, S.F., 2020. Context aware recommendation systems: a review of the state of the art techniques. Comput. Sci. Rev. 37, 100255. https://doi.org/10.1016/j.cosrev.2020.100255
Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., Xie, X., 2023. Large language models understand and can be enhanced by emotional stimuli. arxiv:2307.11760.
Li, X., Lu, Z., Cai, D., Ma, X., Xu, M., 2024a. Large language models on mobile devices: measurements, analysis, and insights. In: Proceedings of the Workshop on Edge and Mobile Foundation Models. ACM, https://doi.org/10.1145/3662006.3662059
Li, Y., Wen, H., Wang, W., Li, X., Yuan, Y., Liu, G., Liu, J., Xu, W., Wang, X., Sun, Y., Kong, R., Wang, Y., Geng, H., Luan, J., Jin, X., Ye, Z., Xiong, G., Zhang, F., Li, X., Xu, M., Li, Z., Li, P., Liu, Y., Zhang, Y.Q., Liu, Y., 2024b. Personal LLM agents: insights and survey about the capability, efficiency and security. arxiv:2401.05459.
Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization arxiv:1711.05101.
Lyu, Q., Apidianaki, M., Callison-Burch, C., 2022. Towards faithful model explanation in NLP: a survey. arxiv:2209.11326.
Mandvikar, S., 2023. Augmenting intelligent document processing (idp) workflows with contemporary large language models (LLMs). Int. J. Comput. Trends Technol. 71, 80–91. https://doi.org/10.14445/22312803/ijctt-v71i10p110
Masoud, M., Jaradat, Y., Manasrah, A., Jannoud, I., 2019. Sensors of smart devices in the internet of everything (ioe) era: big opportunities and massive doubts. J. Sens. 2019 1–26. https://doi.org/10.1155/2019/6514520
Mehrotra, A., Pejovic, V., Musolesi, M., 2021. Futureware: designing a middleware for anticipatory mobile computing. IEEE Trans. Softw. Eng. 47, 2107–2124. https://doi.org/10.1109/tse.2019.2943554
Melumad, S., Pham, M.T., 2020. The smartphone as a pacifying technology. J. Consum. Res. 47, 237–255. https://doi.org/10.1093/jcr/ucaa005
Meta, 2024a. The LLaMA 3 herd of models. arxiv:2407.21783.
Meta, 2024b. LLaMA 3 model card. https://github.com/meta-llama/llama3 (accessed July 2025).
Monge Roffarello, A., De Russis, L., 2022. Understanding and streamlining app switching experiences in mobile interaction. Int. J. Hum.-Comput. Stud. 158, 102735. https://doi.org/10.1016/j.ijhcs.2021.102735
Montesinos López, O.A., Montesinos López, A., Crossa, J., 2022. Overfitting, Model Tuning, and Evaluation of Prediction Performance. Springer International Publishing, pp. 109–139. https://doi.org/10.1007/978-3-030-89010-0
Moustafa, A.A., Diallo, T.M.O., Amoroso, N., Zaki, N., Hassan, M., Alashwal, H., 2018. Applying big data methods to understanding human behavior and health. Front. Comput. Neurosci. 12, https://doi.org/10.3389/fncom.2018.00084
Munappy, A.R., Bosch, J., Olsson, H.H., Arpteg, A., Brinne, B., 2022. Data management for production quality deep learning models: challenges and solutions. J. Syst. Softw. 191, 111359. https://doi.org/10.1016/j.jss.2022.111359
Nakamura, T., 2015. The action of looking at a mobile phone display as nonverbal behavior/communication: a theoretical perspective. Comput. Hum. Behav. 43, 68–75. https://doi.org/10.1016/j.chb.2014.10.042
Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., Seifert, C., 2023. From anecdotal evidence to quantitative evaluation methods: a systematic review on evaluating explainable AI. ACM Comput. Surv. 55, 1–42. https://doi.org/10.1145/3583558
Olawade, D.B., Wada, O.Z., Odetayo, A., David-Olawade, A.C., Asaolu, F., Eberhardt, J., 2024. Enhancing mental health with artificial intelligence: current trends and future prospects. J. Med. Surg. Public Health 3, 100099. https://doi.org/10.1016/j.glmedi.2024.100099
OpenAI, 2023. GPT-4 technical report. arxiv:2303.08774.

Parate, A., Böhmer, M., Chu, D., Ganesan, D., Marlin, B.M., 2013. Practical prediction and prefetch for faster access to applications on mobile phones. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, https://doi.org/10.1145/2493432.2493490

Parthasarathy, V.B., Zafar, A., Khan, A., Shahid, A., 2024. The ultimate guide to fine-tuning LLMs from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities. arxiv:2408.13296.

Patil, R., Gudivada, V., 2024. A review of current trends, techniques, and challenges in large language models (LLMs). Appl. Sci. 14, 2074. https://doi.org/10.3390/app14052074

Patil, H., Mahandule, V., Ajgaonkar, O., 2025. Balancing data privacy ethics in the age of big data:challenges and solutions. J. Innov. Bus. Ind. 3, 1–6. https://doi.org/10.61552/jibi.2025.01.001

Pejovic, V., Lathia, N., Mascolo, C., Musolesi, M., 2015. Mobile-based experience sampling for behaviour research arxiv:1508.03725.

Peng, D., Fu, Z., Wang, J., 2024. Pocketllm: enabling on-device fine-tuning for personalized LLMs. arxiv:2407.01031.

Peters, H., Liu, Y., Barbieri, F., Baten, R.A., Matz, S.C., Bos, M.W., 2024. Context-aware prediction of active and passive user engagement: evidence from a large online social platform. J. Big Data 11, https://doi.org/10.1186/s40537-024-00955-0

P.S. D.V, 2023. How can we Manage biases in artificial intelligence systems – a systematic literature review. Int. J. Inf. Manag. Data Insights 3, 100165. https://doi.org/10.1016/j.jjimei.2023.100165

Rahma, A., Wantini, W., 2024. Human behavior in social context. J. Impresi Indones. 3, 411–417. https://doi.org/10.58344/jii.v3i6.5049

Ram, N., Yang, X., Cho, M.J., Brinberg, M., Muirhead, F., Reeves, B., Robinson, T.N., 2019. Screenomics: a new approach for observing and studying individuals' digital lives. J. Adolesc. Res. 35, 16–50. https://doi.org/10.1177/0743558419883362

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection in machine learning arxiv:1811.12808.

Reeves, B., Ram, N., Robinson, T.N., Cummings, J.J., Giles, C.L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., Gagneja, A., Brinberg, M., Muise, D., Lu, Y., Luo, M., Fitzgerald, A., Yeykelis, L., 2019. Screenomics: a framework to capture and analyze personal life experiences and the ways that technology shapes them. Hum.–Comput. Interact. 36, 150–201. https://doi.org/10.1080/07370024.2019.1578652

Sarker, I.H., Salah, K., 2019. Appspred: predicting context-aware smartphone apps using random forest learning. Internet Things 8, 100106. https://doi.org/10.1016/j.iot.2019.100106

Sağbaş, E.A., Korukoglu, S., Balli, S., 2020. Stress detection via keyboard typing behaviors by using smartphone sensors and machine learning techniques. J. Med. Syst. 44, https://doi.org/10.1007/s10916-020-1530-z

Shapsough, S., Hesham, A., Elkhorazaty, Y., Zualkernan, I.A., Aloul, F., 2016. Emotion recognition using mobile phones. In: 2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE, https://doi.org/10.1109/healthcom.2016.7749470

Singh, C., Inala, J.P., Galley, M., Caruana, R., Gao, J., 2024. Rethinking interpretability in the era of large language models. arxiv:2402.01761.

Solomon, A., Shapira, B., Rokach, L., 2022. Predicting application usage based on latent contextual information. Comput. Commun. 192, 197–209. https://doi.org/10.1016/j.comcom.2022.06.005

Stachl, C., Au, Q., Schoedel, R., Buschek, D., Völkel, S., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., Bühner, M., 2019. Behavioral patterns in smartphone usage predict big five personality traits. PsyArXiv. https://doi.org/10.31234/osf.io/ks4vd

Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., Bühner, M., 2020. Predicting personality from patterns of behavior collected with smartphones. In: Proceedings of the National Academy of Sciences, vol. 117. pp. 17680–17687, https://doi.org/10.1073/pnas.1920484117

Teng, S., D'Alfonso, S., Kostakos, V., 2024a. A tool for capturing smartphone screen text. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, https://doi.org/10.1145/3613904.3642347

Teng, S., Zhang, T., D'Alfonso, S., Kostakos, V., 2024b. Predicting affective states from screen text sentiment. In: Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, https://doi.org/10.1145/3675094.3678489

Tong, H.L., Quiroz, J.C., Kocaballi, A.B., Fat, S.C.M., Dao, K.P., Gehringer, H., Chow, C.K., Laranjo, L., 2021. Personalized mobile technologies for lifestyle behavior change: a systematic review, meta-analysis, and meta-regression. Prev. Med. 148, 106532. https://doi.org/10.1016/j.ypmed.2021.106532

Tonti, S., Marzolini, B., Bulgheroni, M., 2021. Smartphone-based passive sensing for behavioral and physical monitoring in free-life conditions: technical usability study. JMIR Biomed. Eng. 6, e15417. https://doi.org/10.2196/15417

Trifan, A., Oliveira, M., Oliveira, J.L., 2019. Passive sensing of health outcomes through smartphones: systematic review of current solutions and possible limitations. JMIR Mhealth Uhealth 7, e12649. https://doi.org/10.2196/12649

Trull, T.J., Ebner-Priemer, U.W., 2009. Using experience sampling methods/ecological momentary assessment (esm/EMA) in clinical assessment and clinical research: introduction to the special section. Psychol. Assess. 21, 457–462. https://doi.org/10.1037/a0017653

Twenge, J.M., Spitzberg, B.H., Campbell, W.K., 2019. Less in-person social interaction with peers among U.S. adolescents in the 21st century and links to loneliness. J. Soc. Pers. Relatsh. 36, 1892–1913. https://doi.org/10.1177/0265407519836170

van Berkel, N., Ferreira, D., Kostakos, V., 2017. The experience sampling method on mobile devices. ACM Comput. Surv. 50, 1–40. https://doi.org/10.1145/3123988

van Berkel, N., D'Alfonso, S., Susanto, R.K., Ferreira, D., Kostakos, V., 2022. AWARE-light: a smartphone tool for experience sampling and digital phenotyping. Pers. Ubiquitous Comput. 27, 435–445. https://doi.org/10.1007/s00779-022-01697-7

Vega, J., Bell, B.T., Taylor, C., Xie, J., Ng, H., Honary, M., McNaney, R., 2022. Detecting mental health behaviors using mobile interactions: exploratory study focusing on binge eating. JMIR Ment. Health 9, e32146. https://doi.org/10.2196/32146

Wahle, F., Kowatsch, T., Fleisch, E., Rufer, M., Weidt, S., 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. JMIR Mhealth Uhealth 4, e111. https://doi.org/10.2196/mhealth.5960

Wanner, J., Herm, L.V., Heinrich, K., Janiesch, C., 2022. The effect of transparency and trust on intelligent system acceptance: evidence from a user-based study. Electron. Mark. 32, 2079–2102. https://doi.org/10.1007/s12525-022-00593-5

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. arxiv:2201.11903.

Xu, D., Yin, W., Jin, X., Zhang, Y., Wei, S., Xu, M., Liu, X., 2023. Llmcad: fast and scalable on-device large language model inference. arxiv:2309.04255.

Yalcin, E., Bilge, A., 2021. Novel automatic group identification approaches for group recommendation. Expert Syst. Appl. 174, 114709. https://doi.org/10.1016/j.eswa.2021.114709

Yan, T., Chu, D., Ganesan, D., Kansal, A., Liu, J., 2012. Fast app launching for mobile devices using predictive user context. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services. ACM, https://doi.org/10.1145/2307636.2307648

Yue, Z., Zhang, R., Xiao, J., 2023. Social media use, perceived social support, and well-being: evidence from two waves of surveys peri- and post-COVID-19 lockdown. J. Soc. Pers. Relatsh. 41, 1279–1297. https://doi.org/10.1177/02654075231188185

Zhang, B., Yin, C., David, B., Chalon, R., Xiong, Z., 2015. A context-aware mobile system for work-based learning. Comput. Appl. Eng. Educ. 24, 263–276. https://doi.org/10.1002/cae.21704

Zhang, X., Li, W., Chen, X., Lu, S., 2018. Moodexplorer: towards compound emotion detection via smartphone sensing. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 1. pp. 1–30, https://doi.org/10.1145/3161414

Zhang, P., Jung, G., Alikhanov, J., Ahmed, U., Lee, U., 2024a. A reproducible stress prediction pipeline with mobile sensor data. In: Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 8. pp. 1–35, https://doi.org/10.1145/3678578

Zhang, S., Ma, Y., Fang, L., Jia, H., D'Alfonso, S., Kostakos, V., 2024b. Enabling on-device LLMs personalization with smartphone sensing. arxiv:2407.04418.

Zhang, T., Teng, S., Jia, H., D'Alfonso, S., 2024c. Leveraging LLMs to predict affective states via smartphonesensor features. In: Companion of the 2024 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM, https://doi.org/10.1145/3675094.3678420

Zhao, S., Li, S., Ramos, J., Luo, Z., Jiang, Z., Dey, A.K., Pan, G., 2019a. User profiling from their use of smartphone applications: a survey. Pervasive Mob. Comput. 59, 101052. https://doi.org/10.1016/j.pmcj.2019.101052

Zhao, S., Luo, Z., Jiang, Z., Wang, H., Xu, F., Li, S., Yin, J., Pan, G., 2019b. Appusage2vec: modeling smartphone app usage for prediction. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE, https://doi.org/10.1109/icde.2019.00120

Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., Li, C., Dou, Z., Ho, T.Y., Yu, P.S., 2024. Trustworthiness in retrieval-augmented generation systems: a survey. arxiv:2409.10102.

Zubair, U., Khan, M.K., Albashari, M., 2023. Link between excessive social media use and psychiatric disorders. Ann. Med. Surg. 85, 875–878. https://doi.org/10.1097/ms9.0000000000000112