Empowering Diverse Voices: A Scalable Method for Eliciting Micro-Narratives in HCI Health Research.

ANONYMOUS AUTHOR(S)

Engaging with people's lived experiences is foundational for HCI research, especially in designing and evaluating (mental) health interventions. The field has developed a range of methods for eliciting this information (e.g., interviews, probes, surveys), each balancing data richness with participant burden and required resources. This paper introduces a novel narrative elicitation method to empower people to easily articulate 'micro-narratives' emerging from their lived experiences, irrespective of their writing ability or background. Our approach aims to enable at-scale collection of rich, co-created datasets that highlight target population's voices with minimal participant burden, while precisely addressing specific research questions (e.g., understanding individuals' challenges with a digital intervention). To pilot this idea and test its feasibility we: (i) developed an AI-powered prototype; (ii) deployed it in three mixed-methods studies involving over 380 users; and (iii) consulted with established academics as well as C-level staff at (inter)national nonprofits to map out potential applications.

CCS Concepts: • **Human-centered computing** → **User centered design**.

Additional Key Words and Phrases: Human-AI collaboration, methodology, qualitative data collection

ACM Reference Format:

Anonymous Author(s). 2025. Empowering Diverse Voices: A Scalable Method for Eliciting Micro-Narratives in HCI Health Research.. In *Proceedings of Conference on Human Factors in Computing Systems (CHI '25).* ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The ability to collect stories of participants' lived experiences is crucial to the researcher's ability to understand challenges with—and the key opportunities to provide support for—individuals' efforts to manage their physical and mental health. There are many design contexts where HCI researchers might benefit from the ability to collect a large sample of users' stories quickly and with limited effort (for both the participants and the research team). Imagine, for example, trying to understand the range of emotionally difficult experiences that young people struggle with online; or collecting the challenges that patients experience with receiving a behavioural change intervention in primary care across a large city or a state. Answering such questions likely faces several methodological challenges including (i) potential for *high heterogeneity in responses*, especially across cultural, economic, and educational divides (i.e., a large dataset might be necessary before thematic saturation); and (ii) difficulty with *reliably eliciting rich-enough qualitative information* (e.g., due to lack of time, motivation, or willingness to engage) from a sufficiently large proportion of target participants.

In other words, we so far lack approaches that could collect rich qualitative narratives from participants at scale (e.g., >150 participants) without excessive participant burden and excessive costs of conducting the research: (i) interviews or detailed diary studies provide the necessary depth of understanding of participants' experience, but are resource

Manuscript submitted to ACM

intensive and pose a substantial burden on the populations and the research team; (ii) approaches such as cultural probes are often bespoke and are similarly complicated to scale and resource (cf., [21]); and (iii) while questionnaires (including EMAs) can be deployed at scale, it is difficult to capture the depth of users' emotional and other experiences required (cf., [59]).

Proposed design solution. This paper proposes a novel narrative elicitation method—based on human-centred AI interaction design—that could start to address these issues. Conceptually, we draw on prior research in psychology and HCI that uses vignettes (or 'micro-narratives') to convey emotional experiences in ways that are succinct and understandable to others – see for example the use of vignettes in affective science [40], healthcare research [55], decision-making studies [12], AI fairness and accountability[17, 24], or studies of moral development [8]. Such approaches were traditionally used to collect participants' responses to *hypothetical*, researcher-generated scenarios, as a way of standardising the context that the participant reacts to.

Our proposition is to flip this approach: i.e., to explore the possibility of collecting *participant-generated* vignettes, as a way to enable them to articulate and share succinct-but-understandable stories about a specific aspect of their lives (which we call 'micro-narratives' in the rest of the paper). To the best of our knowledge, such an approach has been rarely used previously; in fact, we struggled to identify any major research study that has explored this direction. This is possibly due to the burden that such vignette-creating interaction would pose on the participants in traditional remote collection data methods (cf., the difficulties of answering a simple open-text question, see [65]), as well as the potential need for training the participants to ensure both a consistent micro-narrative format and text clarity necessary for the stories to be understandable, all while retaining the ability to capture participants' own 'voice' and description of their lived experience.

Paper structure and contributions. In the rest of the paper, we proceed in three steps. First, we outline the **conceptual design and implementation** of an AI-powered system to enable participants to articulate 'micro-narratives' of their personal experiences with only limited effort, regardless of their writing competency. The core design innovation involves scaffolding the user's cognitive trajectory needed to develop a vignette-like narrative of their experience (cf., [52]), drawing on a new human-AI collaboration workflow (cf., [64]).

Second, we **examine the feasibility and acceptability** of such a system through a case study of understanding youth difficulties with social media. This case study includes the deployment of a proof-of-concept prototype in a series of 3 mixed-methods studies, all with youth aged 18-20 yo: an initial pilot (N=100), 2-week co-design (N=30), and an experiment (N = 254) that compared the prototype with the use of an analogous open-ended survey question.

Finally, we start to **map out the potential use-cases** of such narrative elicitation methods more broadly, inspired by our informal discussions with 4 C-level staff at major national and international non-profits, as well as 14 established researchers (median citations 13.1k) across a range of disciplines, including clinical psychology, behavioural health, communication studies, implementation science, and HCI.

Findings overview. Our findings demonstrate that our prototype and the general approach to micro-narrative elicitation are feasible and acceptable both to our case-study participants as well as academic/non-profit experts. In particular, the findings from **qualitative user studies** (initial pilot N=100, and 2-week co-design N=30; see Section 4) show that participants perceived the system as both efficient and easy to use, with a majority (90%) reporting it helpful for articulating their experiences. In particular, the process of developing a vignette-based story step-by-step was positively

received. Some participants reported that engaging with the system helped them 'make sense' of their experiences and facilitated personal reflection and understanding.

These findings are further confirmed by the **quantitative experimental study** (N = 254, see Section 6) that compared the AI-supported vignette development by asking participants to share their experience using an open-text survey question (as the closest currently used comparator). A *between-subject comparison* found that the micro-narrative process was significantly more helpful (than the open text form) for articulating the experience, less difficult to respond to, more accurate for capturing the objective features of the situation and that others who read the final product would understand participants' experience better. A *within-subject comparison*[1] found that participants were 4.5x more likely to report that micro-narrative was easier to capture their experience, 2.8x more likely that it was more appropriate for (other) youth, 6x more likely that is it more helpful for making sense of the experience, and 4.8x more likely to better support them in thinking about how to address their social media challenges.

Finally, the **expert engagements** showed a similarly positive response. For example, all four non-profits and seven researchers are currently actively working with us to incorporate the narrative elicitation into their work – both as part of 'pure' data collection, as well as an embedded part of intervention design.

Implications and next steps. Such promising initial findings suggest that this—or other similar—narrative elicitation methods could: (1) Help researchers collect valuable narrative data about clinical and personal experiences at a scale and speed that would be difficult to achieve otherwise, and that they complement existing qualitative approaches like interviews (more qualitative richness but also higher effort required) and questionnaires (lower qualitative richness). (2) Enable participants to actively participate in building and shaping the narrative of their personal experiences in persuasive and understandable ways, e.g., even if they are not confident/skilled writers. (3) Lead to potentially rich, co-created datasets that highlight the voices of the target population, while targeting a specific type of experience/situation being researched.

To support further exploration of this intriguing research and design direction, we describe the full system design—including all LLM prompts—in Appendix A; and will make the system source code publicly available under a Creative Commons license at the point of publication. The code will be further modularized to ease further adaptation and extensions by other research teams. *Note – A version of the full prototype will remain accessible to the reviewers here during the review process.*

2 Related Work

The design challenges raised in the introduction are connected to multiple areas of research, which we review below: We start by outlining the amazing breadth of data collection approaches employed within HCI in the context of collecting participants' lived experiences, with a specific focus on how these balance the trade-offs between qualitative richness in the collected datasets and the participants/researcher effort this richness requires. We then briefly overview the research in psychology around vignettes, which served as inspiration for our approach – i.e., seeing vignettes as a potentially useful form of 'template-driven storytelling'. Finally, we review the recent work in HCI that explored the use of Large Language Models (LLMs) to support digitally mediated data collection as well as narrative (or reflection-oriented) support.

[1]Note that by within-subject comparison, we refer to the fact that participants were making a within-subject judgement between within-subject conditions. However, since participants provided one response (on a bipolar scale), the model was still a between-subjects model.

2.1 HCI methods to elicit / collect participant experiences

Deeply engaging with the lived experiences of stakeholders is a fundamental concern for human computer interaction (HCI) research, with so much excellent methodological and conceptual scholarship published over the last decades that we are unable to fully cover it here (cf., [5, 19, 44, 51]). In the context of this paper, therefore, we specifically focus on the ways in which most commonly used existing methods—such as interviews, diary studies, probes, questionnaires, and ecological momentary assessments (EMA)— *trade-off striving for qualitative richness of the collected datasets, and the required effort from the researchers / participants to do so.*

For example, interviews can be seen as requiring substantial researcher and participant effort to collect (including scheduling coordination, and the time spent talking), in addition to the non-trivial analytical resources required to make sense of the collected data – but result in a deep, nuanced understanding of participants perspectives. Similarly, ethnographic methods might utilise a combination of interviews and long-term participatory observations to seek even more granular understandings about people's beliefs, lived and felt lives [63], as well as in-depth analysis of social practice within their communities; which further increases the resources required from the researcher within data collection & analysis. Diary studies and other probe-based approaches[2] are a related set of qualitative methodologies which seek to address the challenges of developing a longer-term understanding of participants' lived experience. The difference in approach is that these methods do so by deploying digital or physical data-collection instruments directly into participants' daily lives. In the context of our argument here, such methods are thus transferring some of the data collection burden onto the participants (e.g., requiring a daily diary entry, or a thoughtful engagement with a cultural probe) albeit often in playful and reflection or insight-provoking ways – and are intended to inspire ideas and prompt a deeper dialogue between researchers and participants [14, 15, 20]. Finally, methods such as questionnaires and more recently ecological momentary assessments are traditionally built with a different goal in mind – a large-scale collection of mostly quantitative information, often in the form of pre-determined, multiple-choice questions; and with known challenges to reliably collecting open-ended qualitative insights (cf,. [65]). In these cases, the trade-offs tilt towards low-burden (for researchers and, hopefully, participants) which then enables large-scale data collection, but has the side effect of reducing the qualitative richness of information that can be collected.

We note that none of the qualitative methods outlined above (interviews, ethnographies, diary studies, or probes) are traditionally used to collect data from hundreds or more participants in HCI (cf., [6]) – these methods have not been practically or epistemologically designed to answer questions requiring such high sample sizes, at least not for short term, iterative studies. It is, however, not uncommon for multi-year projects in sociology or anthropology to require in-depth interviews with hundreds of participants when complex social questions are explored (cf., for influential examples relevant to HCI [22, 32, 60]) – but the time, financial, and human resource costs necessary for the data collection *and* analysis within such projects are immense.

In summary, the existing approaches do not provide an easily accessible set of methodological tools that would help address the type of questions outlined in the introduction. In other words, at the moment, if one needs to collect qualitatively rich stories from a large number of participants, the existing methods seem to be associated with high burden and cost for either the researchers or participants (and most commonly both) – even if there are examples of research questions for which such a price is worth paying.

[2]We note that the flexibility and wide-spread interest in probes has meant that interaction design researchers have taken on and adapted probes in many ways, leading towards more varied applications as well as epistemological and conceptual disagreements on what the role / form of probes could be – see for example [3, 68]. For the purposes of our argument here, we will point the reader to the literature that differentiates among, for example cultural probes [14], technology probes [20], empathy probes [37] and informational probes [9].

2.2 Vignettes as examples of usefulness of short narratives in psychology research

In preparation for the conceptual system design in Section 3, we now turn our attention to 'vignette' studies. This methodology, initially emerging from ethnographic research, uses *vignettes*—i.e., short scenarios—as a way of providing participants with succinct and widely understandable representations of a situation of interest, which the participants then react to. For the purpose of this paper, our interest in vignettes is not in how they are currently used—as a prompt to elicit a participant response—but rather in what these studies implicitly show about the vignette format *as a 'storytelling device'*: that is, their apparent ability to convey potentially intricate and emotional stories within a short and template-like 'micro-narrative'.

The rest of the section will briefly outline the various ways vignettes have been used in social sciences, with the aim of illustrating what is already known about vignettes as 'structured-storytelling;' that is, as a form that encapsulates stories and emotional experiences (albeit, to date, primarily stories crafted by researchers). This is preparation for Section 3, where our main design question will be focused on how we could rethink the communication purpose of vignettes from conveying stories from researcher-to-participant, and instead utilise the vignette's form factor as a scaffolding to enable the participants themselves to express their own stories.

Vignettes as qualitative response elicitation. Much of research in social sciences has focused on using vignettes as 'inputs' into a qualitative interview process in social research, helping the researchers unpack topics that might be otherwise difficult to engage with and struggle with social desirability bias. In this sense, the short 'descriptive scenarios' can effectively elicit responses to sensitive topics, and enable the participants to focus on elements that are particularly important for the researchers' topics of interest – with prior use both within sense-making work as well as intervention development (cf., [56]).

In these instances, the researchers place importance on creating believable and 'realistic' vignettes to reduce the tendency of participants to answer in general / abstracted / hypothetical terms. For example, Sampson et al [45] articulate the value of their use of 'real-life' (i.e., directly based on field observations / interviews) with the following quote:

> *Overall contribution of 'real-life' vignettes to the outcomes of studies A and B, we consider that their greatest impact was in encouraging participants to engage with the materials presented to them to such an extent that interviewers were temporarily granted insider status within their 'communities of practice' [29]. Here the vignettes: stimulated engagement and openness; reduced the tendency for idealised answers; facilitated the development of a high degree of trust in situations where participants were suspicious; and generated credibility. This allowed participants to discuss matters that would generally be off-limits. In this context, they were able to reveal the 'unacceptable' (errors, deviant/prohibited practice, non-masculine behaviour) and reflect on the proscribed.*

The use of vignettes in similar contexts is thus exploratory and/or interpretative: aiming to help elicit in-depth qualitative responses from participants to better understand the 'tacit' knowledge and social practice that might be otherwise difficult to uncover; whilst offering participants carefully scoped 'microcosms' (cf., [54, p. 343]) to react to. The use of vignettes can accomplish these goals by providing participants with examples of real situations and experiences they can identify with, which communicates to participants the researcher's status as a quasi-insider, lowering barriers to open communication.

Vignettes as an experimental research tool. The second way in which vignettes are used in research is as tools in experimental research. Such an approach is commonly applied in experimental psychology research, such as fields

examining decision-making, moral judgments, or cognitive psychology. In these instances, the focus is on uncovering the generalisable cognitive patterns (or their variance across populations). The (sets of) vignettes are seen as providing a stable and easily-understandable set of stimuli, which enable the researcher to manipulate and examine the impact of varying parameters of importance. For example, [49] describes using such an approach to understand the inequalities derived from 'unwarranted variations' in health care — such as those hypothesised due to implicit biases (e.g., delays in cancer or poorer reported experiences with doctors for patients of marginalised backgrounds).

In these contexts, sets of vignettes are created to share a common structure while allowing for variation in elements that are assumed to have theoretical importance[3]. In this instance, vignettes are described as:

> *short, carefully constructed depiction of a person, object, or situation, "representing a systematic combination of characteristics" […] In experimental vignette studies, vignettes are used to explore participants' attitudes, judgements, beliefs, emotions, knowledge or likely behaviours by presenting a series of hypothetical yet realistic scenarios across which key variables have been intentionally modified whilst the remaining content of the vignette is kept constant. Such studies seek to generate inferences about cause-and-effect relationships by considering the nature of each vignette, and participants' subsequent responses to these vignettes.* [49]

We note that vignettes are still serving a communicative role in these studies and are often designed to convey / elicit emotions. However, in contrast to the qualitative studies, more focus is given to a theory-informed template-like form of the vignette. Interestingly, this then allows individual aspects of the vignette-described story to be independently manipulated while retaining story coherence – an observation that we will use in our design in section 3.

2.3 Digitally-mediated data collection and narrative/reflection support

Finally, there has been an explosion of interest in recent years in understanding the range of human-AI collaboration tasks, especially as enabled by the rise of generative AI and Large Language Models. In the context of this work, we are interested in two streams of this research: the first is exploring the support of *participants' reflection or articulation*, which is to date mostly centred under the banner of creativity; the second is the focus on *streamlining data collection* techniques. We review the most relevant work from each of these streams below:

Reflection & articulation. Much recent work has focused on the role that LLM systems can play in supporting a spectrum of human-AI co-creation tasks (cf., [36]). For example, Luminate [31] is designed to help users explore and navigate a wide range of possible ideas. The LLM-components of the system are used to (i) generate possible categories; and then (ii) allow the user control and exploration of the range of options with a 'dimension-guided' response generation. As a further example of using LLM systems for creative tasks, the 'Idea Machine' was developed to facilitate idea generation and reflection, enabling users to expand upon, rewrite and connect ideas [11]. Research has also examined the use of LLMs to edit written documents, demonstrating how these systems can effectively address both grammatical and content changes [27]. Recent studies in education have also considered the potential of LLMs to provide personalised feedback on students' written work, finding that these tools enhance students' writing proficiency and motivation [38, 41]. Across all of these examples, the research suggests that there is an opportunity for LLMs to empower human creativity, including articulation of narratives – while still enabling the user to remain in charge and keep control over the final outputs.

[3]We note that such use of systematically varied vignettes is also increasingly common in HCI, e.g., in the areas of understanding trustworthiness & transparency of AI decision making: cf., for example [2] for a well cited instance, examining the impact AI-decision explanation styles (input, sensitivity, case-based, demographics) have on perceived fairness of the resulting decision.

Streamlining data collection – surveys. Recent research started exploring the potential of AI-powered interactions as an alternative form of data collection from users. For example, [65] investigated the potential of pre-GPT chatbot systems to conduct conversational surveys with open-ended questions. The key motivation was to reduce survey fatigue and user burden, with the aim to reduce the likelihood that users will skip such questions or provide low-quality responses. Results from their experimental study showed how the chatbot driven surveys had substantially higher completion rate, and somewhat higher informativeness and relevance (with much fewer 'gibberish' answers), with participants volunteering more detail and engaging for longer. These findings were motivated by prior research suggesting that engaging chatbot systems as data collection tools may improve response quality, participant engagement, and enjoyment relative to traditional survey methods [1, 25]. These chatbot systems may be further enhanced by humanization techniques, which improve respondents' perceptions of the chatbot and increase interaction time [43].

Streamlining data collection – narratives. In the last months, several research projects have extended such an approach from survey open question to exploring the potential of Large Language Models in collecting of longer form user narratives, often in the form of a chatbot-led 'interview': Wei at al. [59] developed a set of GPT-3 powered chatbots to "collect user self-reports while carrying naturalistic conversations", motivated by the challenges with tracking burden for participants if such details were to be captured manually. Focused on four health context (sleep, food, work productivity, and exercise), the authors showed how carefully prompt-engineered bots were able to collect data on pre-determined aspects of the context in question (e.g., food intake for each of the main meals of the day), although not in an entirely reliable way. The authors also highlight the ethical & design challenges of potential problematic responses – cf., the wider discussion on LLM-alignment [13].

In more clinically sensitive work, Kim [26] developed a "MindfulDiary" to support psychiatric patients' journalling experiences, with the interaction driven by AI-generated prompts and reflective questions (driven by OpenAI's GPT4). The authors aim was to reduce the burden that traditional daily diary entries would impose on patients, as well as enhance self-exploration and aid in expressing their experiences and emotions. The resulting diary interactions were well received by patients and therapists, with the system helping to mitigate some of the challenges associated with traditional journalling approaches: resulting in interactions that "ensured the users are not overwhelmed by the task, and guided in documenting their feelings and experiences more richly." Finally, Seo et al [47] developed a LLM-powered system to empower children to share and reflect on their emotional experiences. The short-term lab testing with 20 children showed promising results, with children being receptive to using ChaCha for disclosing emotions and stories – even, concerningly, those that they have never shared with their parents.

These systems show the promise of LLMs to help reduce participant burden and enable innovative approaches to supporting data collection, including the focus on supporting users' narrative creation. However, most of these examples still shared the challenge of building reliable and robust systems based on prompt engineering inherently stochastic Large Language Models – cf., also [67], and were often designed for highly specialised use contexts.

3 Part 1: Developing the prototype

The proposed data collection framework initially started as our attempts to answer a specific question: *"How can we enable young adults to share their narratives of the specific challenges that they face with social relationships as mediated by*

their use of social media?". We were interested in collecting young people's own detailed stories of what has happened online for them, and how they then interpreted/reacted to these situations[4].

Our initial vision was to develop a system that would enable us to collect 2-3 such stories from 200-300 youths across the country and socioeconomic backgrounds. We hoped to thematically analyse the resulting dataset of up to 900 stories to understand the common themes of what is seen as difficult by the youths themselves ('canonical stories') as well as how such canonical exemplars vary across the country / socioeconomic divides. We expected the results to lead to both additional *qualitative work* (e.g., interviews/co-design using the exemplar stories across themes as prompts to more deeply understand youths' lived experiences and to co-produce possible digital interventions to address these), as well as further *quantitative large-scale confirmatory research* across dozens of schools based on our pre-existing collaboration with a government-supported non-profit organisation.

It was only after we developed the first prototype for this specific question and research plan—and faced an unexpected interest in other applications from our colleagues across highly different projects and use-cases—that we realised how the underlying interaction design could be abstracted and conceptualised as a much broader **research method** for elicitation of participants qualitative experiences. This insight then led to both the expert engagements described in section 7, as well as the conceptualisation of the design space that we move onto now.

3.1 Design conceptualisation

In the rest of the paper, therefore, we propose that the design questions we were trying to resolve for the 'youth social media stories context' can be seen as just one instance of much broader design aims. We articulate these as based around the design tensions of requiring a remote narrative elicitation method that, at the same time, provides: (1) enough *open-endedness* (to capture participants' lived experience, in their words); (2) enough *consistency* in the set of core aspects covered within the stories (so research questions such as the one above can be addressed); while (3) *reducing the burden* for participants (so they remain willing to create & share their narratives, and ideally find value for themselves in doing so).

Drawing on the prior work concerning user-burden often associated with remote open-text data collection [39, 65], we assumed that the key design innovation will need to be in reducing the associated friction and cognitive difficulty inherent to text-based narrative creation especially for marginalised / at-risk populations (cf., [16, 26, 47]. As such, we were interested in exploring how the system could reduce user burden by including some form of scaffolding for the articulation process (cf., Section 2.3).

Design approach. The core insight driving our design answers to this dilemma of openness vs consistency vs burden was the way in which vignettes—as a narrative structure—seem to support a form of succinct-but-understandable storytelling. As we outlined earlier in Section 2.2, many vignettes in prior work can be seen as instantiations of particular 'templates' that provide focus for the resulting story: these do not constrain the content of the situation being described, but guide the author by specifying a set of aspects of the situation that should be included.

We wondered if—using the same line of thought—it *might be possible to elicit stories describing youth challenging experiences with social media as a consistent template*: for example, by combining answers to pre-determined elements describing the 'specific event that happened', any 'additional context that is necessary to understand the event', the

[4]We note that this is a surprisingly under-studied question in the social communication field: despite the extensive research examining social media's wellbeing impacts, the specific social media challenges impacting young people's wellbeing remain poorly understood (cf., [28, 57]). Moreover, within social media research, there continues to be a lack of qualitative research that considers young people's lived, individual experiences to answer questions about social media's wellbeing harms (see also [23, 46, 60]). This research gap is compounded by an absence of youth consultation within existing studies, which fail to adequately capture youth voices and/or experiences of social media harms(e.g., [58]).

'resulting emotional reaction', and what the young person 'found worst about the situation'. We were further curious to see whether such a template would be seen as overly restrictive, or in fact serve as a welcome articulation support – e.g., as set of 'guide rails' that provide the participants with enough freedom to choose how they describe the key elements of the situation, while offering them a cognitive structure of what aspects to consider.

Design assumptions. Based on the considerations above, we have articulated three design assumptions, that drove both our development work (described below) as well as the empirical studies (described in the next sections). We outline these below:

DA-1 The first design assumption we wanted to explore was whether supporting articulation of a story might consist of *collecting a set of carefully selected 'fragments' of the situation* (which are individually easy to answer), but which can be then *combined into a coherent narrative* (based on an underlying template) and serve as a helpful starting point for further adaptation if needed.

DA-2 Our second design assumption was that *this process could be supported through Large-Language Models*, targeting the steps in the cognitive process above that might be particularly contributing to the user burden. In particular, we wanted to explore the potential of 'agentic' human-AI workflows: i.e., approaches where the target task is broken down into individual components, where 'specialised' LLM components are used to support individual steps of tasks, at least some of the programme execution are governed by LLM outputs.

DA-3 Finally, our third design assumption was that, if we are successful, such articulation support process will be perceived by the participants as:

- *Simple and efficient*: e.g., easier than composing a narrative from scratch, such as in a open-text survey;
- *Empowering and affirming*: enabling articulation of an experience in ways that feel to be "the user's own voice" even if the user received LLM support; and potentially also
- *Personally helpful / insightful* – e.g., that the process of articulation might lead to new insight / perspective on the reported experience.

In summary, the first assumption aimed to describe the 'theory-of-change' that would underpin the proposed new capability (cf., [52]) – i.e., guide how and why we would believe the interaction design would lead to the intended outcome. The second assumptions highlighted the technical innovation which we assumed could not only reduce user-burden but actively enhance their ability to express themselves (and the novelty of which could explain why a similar design has not yet been proposed). Finally, the third assumption then focused on the user-implications of the resulting process – each aligned with the goals of reducing user-burden, facilitating open-ended self-reflection, and increasing perceived benefit.

3.2 Proof-of-concept implementation

Two complementary design frameworks drove our design exploration and attempts to test the assumptions above.

Designing for cognitive trajectories. First, we draw on Slovak & Munson framework [52] and its focus on clearly articulating the 'cognitive trajectory' that the design is aiming to support. In this case, we considered how the mental process of accomplishing this task might look like for the participants if they were to be asked to articulate similar template-based narratives without any AI-support; and, most importantly, *which of the steps in such a flow might be most difficult and/or burdensome*. For example, a plausible expectation[5] would be that participants might be first asked

[5]As an example, the Institute for Sustainable Futures toolkit includes a similar activity flow, which is intended to be administered as part of the facilitator-led workshops – see link here.

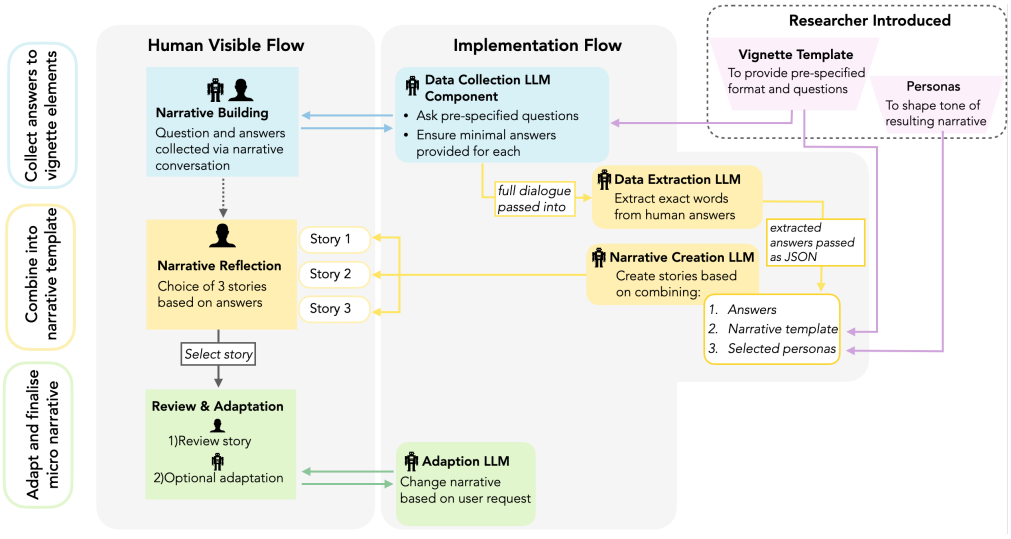

Fig. 1. Overview of the Three Stage Human AI Workflow.

to *consider responses* to the *individual questions* within the template, then *reflect on the connections* between answers, and finally *combine and rewrite* their answers into a narrative that incorporates all of the information (potentially with adding further detail or editing previous answers). Our preliminary assumption was that the most difficult steps would likely be the *combine and rewrite stage*, especially if the proposed template is not closely aligned with a mental model that participants would have.

LLM-chaining as 'crowdsourcing'. Second, we were inspired by the recent approaches based on LLM-chaining as an analogy to prior crowdsourced workflows [64]. In our view, this appeared to be a promising direction that would enable the need to maintain fixed (state-based) trajectory of the cognitive flow, with the openness and flexibility of LLMs for individual steps (cf., [26, 47, 50, 59]). In addition, such an approach allowed us to 'enforce' the assumed theory-of-change (i.e., the cognitive trajectory that we aimed to support), while the identified human/AI subtasks encapsulate the AI prompting into independent and easier-to-debug chunks [64]). Finally, we also assumed that such infrastructure adds useful modularity to the resulting human-AI workflow (cf., [50, 67]), e.g., allowing inclusion targeted safeguarding components (such as detection of self-harm / suicidality risk) into the workflow without the need to adapt other steps.

Across all of these, our implementation was based on prompt engineering literature (e.g., [61]) as well as our prior expertise with similar prompt engineering and design projects. We selected the underlying frameworks (langchain & streamlit) to enable rapid exploration and change, with the view that this will enable deeper co-production and active engagement from our youth participants (as well as experts) – which we assumed will be a necessary component given

the focus on developing a new human-AI cognitive workflow, without clear prior literature on neither the expected human-AI distribution of 'tasks' in support of story articulation nor an understanding of user preferences.

3.2.1 Proposed Human-AI workflow flow . The envisioned workflow consisted of three stages described below and visualised in Figure 1. The aim was to mimic the cognitive flow outlined above, and retain participants' ability to share their stories in their own words ('voice') in as much as possible while reducing likely friction through LLM support in as much as possible.

> *Stage 1 – Data collection:* Human-AI collaboration to collect answers to the selected questions (i.e., story 'fragments' that form the elements of the vignette) through a conversational interface – the LLM is prompt engineered to ensure that the participant is asked—and provides answers—to all questions specified in the vignette template through natural language conversation (cf., [47, 59]).
>
> *Stage 2 – Synthesis and narrative building:* Once the questions are answered, an AI-only component extracts the exact answers that users provided, and then combines these into a proposed narrative following the vignette template that (i) leads to a coherent narrative; (ii) directly incorporates what users' said in their own words; while (iii) attempting to add as little additional information as possible. This narrative-building step is repeated three times, each time with a different 'persona' that aims to affect the tone of the 'connective text' that the LLM inserts when connecting the user's answers to the 'fragments' into a coherent story.
>
> *Stage 3 – Review and selection:* Finally, users review the three versions of their story, and are asked to see if any of these 1) resonates with the 'content' of what they wanted to express and 2) feels like 'their own' voice (Human-only choice). We chose to provide three different 'tones' of the story—implemented as different 'personas' behind the scenes—to encourage a feeling of meaningful choice of which narrative feels best like 'their own' and also highlight the range of ways in which their words can be combined together. The participants also have the opportunity to further adapt the story as needed, e.g., if the LLM workflow misrepresented any information, or did not include something the participants find crucial to their experience (Human-AI collaboration).

3.2.2 Implementation details. The implementation for each of these stages is described in detail within Appendix A, including the full prompts used. The source code will be published on github under a Creative Commons license and made freely available upon publication; for the review process, we attached the source code as a supplementary file on PCS.

4 Overview of the design & empirical studies

To test the feasibility and acceptability of the AI-driven prototype system outlined in Part 1, we carried out two streams of empirical work. In the first, we conducted three mixed-method studies that considered the prototype within the specific context of social media challenges experienced by youth. In doing so, we aimed to test the design assumptions outlined in Part 1 within a relevant case study and consider how this novel data collection process compares to traditional survey methods, from a participant perspective. These studies were designed sequentially such that the output from each informed the design and implementation of the next study, allowing us to holistically examine the different aspects of the interaction of relevance to our design assumptions. In parallel, we engaged both expert researchers and non-profits in informal engagements to probe how this prototype, and the specific process it involved, may apply to other case studies or research contexts (stream 2). The aim of this component was therefore to loosely map out the potential

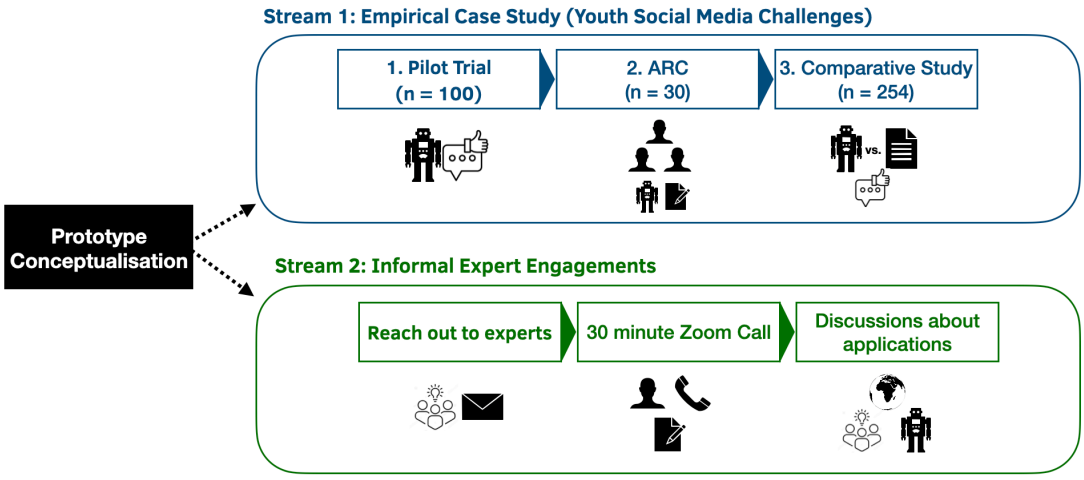

Fig. 2. Overview of the Research Components

application of this new data collection method within a broader research/health context. These complimentary research processes are outlined in Figure 2, with an overview of how each maps onto the design assumptions in Figure 3.

5 Part 2: Initial empirical studies

To test our design assumptions, we conducted two technology probe online trials: a Prototype Pilot Study and an Asynchronous Remote Community (ARC) study. Together these studies allowed us to examine the design aims within the specific case study of young people's social media challenges. Both studies were approved by the institutional ethics committee at [masked for review]. As our prototype employed fairly nascent LLM models, we were cautious about not collecting participants' real data in the initial empirical trials. Instead, we asked them to reflect on *hypothetical* social media experiences. This was an ethical decision we made as we did not yet know exactly how these interactions would play out and we wanted to lower the risk of any sensitive disclosures while we were developing the prototype. However, given that the ultimate goal was to collect participants' *real* experiences, we did so in the final empirical study (Study 3). We made this decision only once we received positive results from the pilot trial and ARC co-design activities, which made us confident that we could safely ask for participants' real experiences in the final experimental study.

5.1 Participants and Procedure

5.1.1 Prototype Pilot Study. The pilot study aimed to examine whether the prototype effectively supported the narrative-building process outlined in the design conceptualisation and understand the user experience of this process. We were also interested in how participants perceived the chatbot as a novel LLM-system. To this end, study 1 was largely focused on testing design assumptions 1 and 2 and capturing general feedback on the prototype/interaction.

100 UK-based participants were recruited via the participant platform Prolific. The decision to use Prolific was informed by our aim to rapidly trial the prototype with a diverse sample of young people across the UK, although we recognise that the demographic may be somewhat skewed reflecting the characteristics of Prolific's participant pool. Based on the available demographic data, 46 participants identified as female, 53 as male, and 1 unspecified, with all participants aged between 18-20 years old. 56 participants identified as White, 32 as Asian, 6 as Black, 4 as mixed, 1 as

	Quantitative and qualitative interaction reflections	ARC - Persona Reflections (Task 1 and 2)	ARC - Interview Discussions (Task 3)	ARC - Application Reflections (Task 4)	Comparison of chatbot and survey interactions
Study	Pilot (Study 1)	ARC (Study 2)	ARC (Study 2)	ARC (Study 2)	Comparison (Study 3)
Assumption 1: *Understanding articulation as a narrative building process*	X	X	X	X	
Assumption 2: *Exploring the potential of LLMs + agentic interactions to support articulation*	X		X	X	X
Assumption 3: *Understanding participant perceptions and perceived benefits of the interaction*					X
General Prototype Feedback	X	X	X	X	X

Fig. 3. Outlining the three studies in relation to the three key design assumptions outlined in Part 1. A grey cell indicates that the study was designed to address the design assumption in the relative row.

Other and 1 unspecified. 32 worked part-time, 32 were unemployed or not in paid work, 31 unspecified or due to start a new job and 5 were full-time. The study consisted of three components: demographics and background questions (Qualtrics), prototype interaction (Streamlit), and feedback survey (Qualtrics), which captured both quantitative and qualitative data about the interaction experience. The prototype interaction is shown in Figure 4. Participants were compensated £10/hour via the recruitment platform (averaging £3.30 for the 20-minute study).

5.1.2 Asynchronous Remote Communities. Having received largely positive feedback about the chatbot interaction in the pilot study, we developed the ARC study to gather richer qualitative reflections that would help us further unpack the articulation process (DA1) and understand how and why participants perceived this as beneficial (DA3). A second aim was to understand how specific aspects of the chatbot (e.g., personas, voice) could be changed to make the interaction more youth-friendly and empowering (DA2, DA3). In this sense, the ARC was designed to triangulate similar design questions to the pilot study, with the added element of enabling participants to co-create elements of the prototype, such as the chatbot's tone and personas This study also generated much richer qualitative data around the articulation process as afforded by in-depth discussions and reflections from participants. We recruited 30 UK-based participants via the participant platform Prolific. We chose participants to represent a range of backgrounds within the UK, to reflect a representative group of young people. Based on the available demographic data, 12 participants identified as female, 18 as male, with all participants aged between 18-20 years old. 15 participants identified as Asian, 9 as White, 4 as Black, and 2 as mixed. 15 were unemployed or not in paid work, 7 unspecified or due to start a new job, 6 worked part-time, and 2 worked full-time. Asynchronous Remote Communities (ARC) are a research method similar to web-based focus groups that leverages digital platforms to facilitate asynchronous discussions and activities among participants(cf.,[34, 35]). Their flexible format makes ARCs particularly suitable to explore and develop new technology

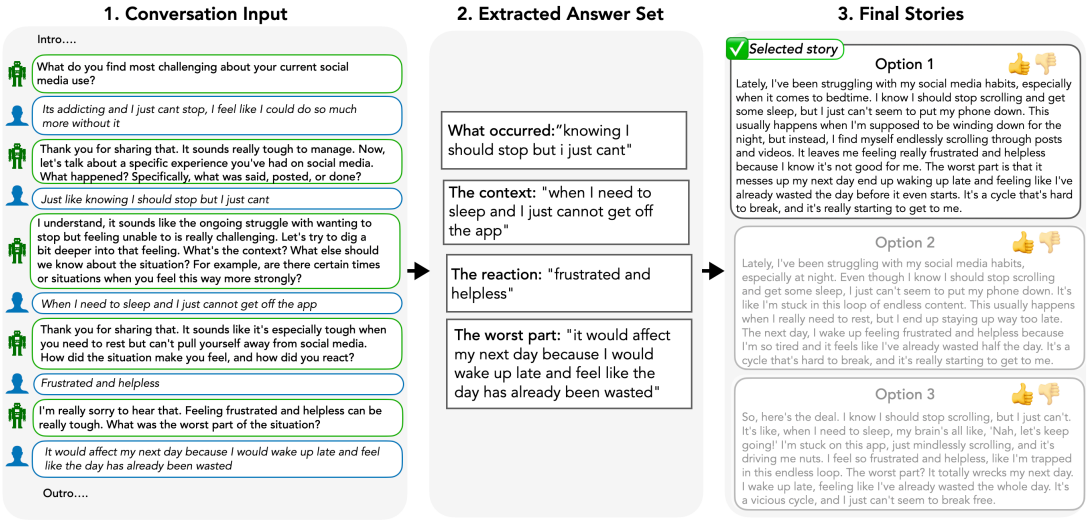

Fig. 4. Example Prototype Interaction from the Pilot Study

probes using co-design activities. For the present study, we chose the Miro platform to carry out the ARCs as it easily facilitates asynchronous discussions and brainstorming activities and supports a range of different co-design tasks that can be accessed across multiple devices (e.g., via computers, tablets, or mobile devices) [7]. Study activities are outlined in Figure 5. We designed activities sequentially, such that each activity was informed by the data collected in the previous activity. This allowed us to iteratively explore participants' experiences and revisit our design assumptions throughout the study.

5.1.3 Data Analysis. Three authors (including the first and last author) met routinely throughout the Pilot and ARC studies to discuss participant feedback, both from the Pilot surveys and ARC activities. For the pilot study, we conducted a descriptive analysis of the quantitative survey data alongside a qualitative analysis of the open-ended survey questions, which were annotated by authors to extract key themes across participant responses. For the ARC study, the first author analysed participant contributions across the four ARC activities using a hybrid thematic coding approach. The approach involved using 'sticky notes' on Miro to visualise and draw linkages between the codes within and across participants from each activity to form themes. The development of the coding scheme was an iterative process and any discrepancies were discussed with Authors 3 and 10 until a consensus was reached. For interviews, data was audio recorded with permission from participants, and the recordings were stored in the University's protected server. All interview data were anonymised before being transcribed and coded using thematic analysis. Finally, a similar hybrid coding approach using Miro was used to derive insights relevant to the design questions *across* both studies.

5.2 Results

We start by providing a general overview of the prototype interactions and how participants perceived these across the studies. Given that both Study 1 and 2 effectively triangulate the same design questions from different angles, we

#	Name	Activity	Format
A1	Trying Out Different Storytelling Voices	Using an internet browser, participants interacted with a web-based LLM chatbot to explore hypothetical social media experiences. They were given the opportunity to explore different personas (e.g., younger sibling, friend) and choose the personas they preferred to articulate their experiences. Following the interaction, they provided feedback via a Qualtrics survey.	Streamlit (LLM Chatbot Interaction) & Qualtrics Survey
A2	Finding the Right Chatbot Voice	Participants accessed a web-based group Miro board to engage in short tasks alongside other participants (n = 15 per group). They provided feedback on several personas explored in Activity 1. They were also asked to reflect on the chatbot's tone, language and personality and design an ideal chatbot persona for young people.	Miro Board
A3	Diving Deeper: Interviews	Participants participated in 30-minute, one-on-one interview with a researcher on Zoom. The interviews probed individual feedback and experiences of the prototype, the process of articulating social media experiences and further applications of the chatbot.	Zoom Interview
A4	Looking Further: Exploring Next Steps and Applications	Participants accessed a web-based group Miro board to engage in short tasks alongside other participants (n = 15 per group). They were asked to brainstorm additional features that could be added to the chatbot to make it more helpful. They also reflected on the pros on cons of using chatbots and explored further potential applications for young people.	Miro Board

Fig. 5. Summary of ARC Activities

combine the findings from both studies here, using these to revisit our main design assumptions. Section 5.2.5 then steps back to review participant feedback for the interaction.

5.2.1 General Overview. Across all studies, participants engaged with the prototype effectively, completing the interaction to generate a range of social media experiences (total n = 130, total micro-narratives generated = 130). The majority of participants (84%) found the interaction valuable. Positive feedback highlighted the chatbot's efficiency, helpfulness and ease of use, with one participant remarking, *"I'm getting paid for my time but I'm also getting some advice from it, so it's really helpful"* (P4). Participants also found the interaction engaging, describing it as both *"entertaining and rewarding"* and highlighting the chatbot's ability to *"understand what young people are going through"* (P6) For the ARC activities, all participants completed the initial interaction (n=30), although we observed some attrition for the remaining activities (Miro Task 1: n = 29, Interviews: n = 10, Miro Task 2: n = 18). We note that the interviews were particularly difficult to recruit for and required several recruitment rounds and additional financial incentives, still yielding only a third of the initial sample.

5.2.2 Assumption 1 - Understanding articulation as a narrative-building process. The majority of participants (90%) found the chatbot helpful in articulating their social media experiences and reported that the chatbot did a good job of articulating these experiences.

Participants described the powerful effect of engaging with their experiences as micro-narratives, which led to emotional realisations or *"epiphanies"*. One participant described this as an immediate realisation, *"After reading the scenario I realised that that is what I was thinking but prior to reading it I hadn't really realised it"* (P9). Another participant characterised this interaction as a *"sense-making"* process, where the act of seeing their experience echoed back to them, leading them to make sense of what had happened. Participants perceived this process as empowering, as it allowed them to voice their experiences in an *"innovative"* way and helped them to *"find the right words"* to articulate themselves. Some also described it as *"therapeutic"* as it allowed them to better understand their feelings and experiences. While participants saw this interaction as valuable for articulating *"base and medium"* problems, they expressed some

Persona	LLM Prompt	Feedback (+/–)
Friend	You're a 23-year-old who is collecting stories of difficult experiences that your friends have on social media. You're trying to use the same tone and language as your friend has done, but you can reframe what they are saying a little to make it more understandable to others.	+ *"It felt the most down-to-earth and realistic. It felt like something a friend would genuinely confide in you to discuss."* – *"Whilst it was really good, it may have come across a bit too formally. However, this could be fine still as being too informal comes across as cringe."*
Younger Sibling	You're a 14-year-old teenager who is collecting stories of difficult experiences that your friends have on social media. Use a language that you assume the friend would use themselves, based on their response. Be empathic, but remain descriptive.	+ *"It was the most accurate that the chatbot presented to me."* – *"Seemed too formal, didn't speak with enough slang to emulate a 14-year-old."*
Influencer	You're a 25-year-old social media influencer who is collecting stories of difficult experiences that your followers have on social media. Use a language that is trendy and engaging, as you would on your social media platforms. Be empathetic, but remain descriptive and relatable.	+ *"It was most accurate to how I think the influencers I consume on social media would speak."* – *"It seemed very superficial. It also seemed stereotypical. I think the problem was the assumption that an influencer must be snobby."*
Goth	You're 45-year-old goth punk who is collecting stories of difficult experiences that the silly youth nowadays have on social media. Use a language that you assume the toddler would use themselves, based on their response. Be edgy and cheeky in your response but remain marginally respectful.	+ *"It was less robotic and more human as opposed to the others. I could imagine the person more clearly in the text."* – *"I did not like the tone of voice used to convey the message, which almost turned the situation into a joke."*
Psychologist	You're an expert developmental psychologist who is collecting stories of difficult experiences that your clients have on social media. Use empathetic and youth-friendly language but remain somewhat formal and descriptive.	+ *"I think it came across formally but in a non-patronising way. It seemed like an adult you would want to listen to and respect their advice."* – *"All it did to emulate this persona was elevate the literacy level and vocabulary used which I feel does not really present a psychologist persona but rather someone of a higher education and/or intellect."*

Fig. 6. Summary of Persona Feedback

concerns about the chatbot being able to articulate deep personal issues, which may require more nuance and personal understanding.

To understand how the chatbot's persona influenced participants' articulation experience, we had them trial different chatbot personas. Of the five sample personas, the 'friend' and 'influencer' were the most popular, likely because their ages (23 and 25) and tone were closest to those of the participants. Overall, these personas were described as relatable and realistic and deemed as important for ensuring the chatbot accurately captured participants' tone in the interaction. For example, one participant described how relatability influenced their perception of the chatbot's ability to articulate their experience, *"I related to this style much more, and felt as though it was my own thoughts being put on paper."* Participants also noted that these personas distinguished the chatbot from other chatbots they had used (e.g., ChatGPT) which felt more rigid and impersonal. Participants also identified the importance of appropriate slang and colloquial language, commenting that the chatbot should incorporate more of these to effectively emulate young people's voices and authentically articulate their experiences. This feedback is summarised in Figure 6

5.2.3 *Assumption 2 - Exploring the potential of LLMs and agentic interactions to support articulation .*

The Potential of Agentic Interactions: Participants seemed to particularly value the agentic nature of the interaction, where the chatbot's systematic, step-by-step questioning process helped them break down their experiences into manageable parts. One participant likened this to an exam process, where *"you go through this part first, explore one feeling, and then move onto the next"* (P2). This structured approach not only made the process of reflection easier but also helped participants to organise or *"streamline"* their thoughts more effectively. The specific and targeted nature of the chatbot's questions was also seen as beneficial, encouraging a deeper and more focused exploration of participants' experiences than would come of less structured conversations. Because of this interaction style, the chatbot was perceived as a more productive way to reflect on an experience than speaking with friends as it got to the

point and enabled more efficient reflections. As one participant noted, *"It made me put my thoughts together a lot better than speaking to a friend"* (P10). In reflecting on the chatbot's unique characteristics, participants highlighted the high degree of impartiality which meant they felt like they could *"speak freely"* about their experiences without worrying about the consequences, with one participant saying that the prototype was *"much less daunting to approach, even than an anonymous hot-line"*.

The Potential of LLM Systems: Several participants expressed a preference for interacting with the chatbot, as an AI system, over people because its unique characteristics (e.g., impartiality, directness) made it more approachable for sharing personal challenges. One participant reflected on this in relation to social rejection, *"Speaking to people you are not able to voice your thoughts as maybe they will reject those thoughts or react harshly but with the chatbot you don't really have that fear... You are able to speak a lot more freely."* (P9) When reflecting on the chatbot's general benefits, participants also mentioned its accessibility, which meant its interactions were not limited to certain contexts (as opposed to conversations with friends or family members). However, a few participants also highlighted the limitations of the chatbot, and AI chatbots in general, compared to human interactions, which may lack emotional depth and involve an absence of trust relative to human interactions.

Further Applications of LLM-Driven Agentic Interactions: To generate deeper reflections on the overall interaction process, we also asked participants to reflect on other situations where the chatbot's perceived articulation process (i.e., systematically breaking down personal experiences) may be helpful (see Figure 7). In doing so, we aimed to test both the perceived benefits of the articulation process (assumption 3) *and* the potential of such LLM-driven agentic interactions to address other problems faced by young people (assumption 2). One participant mentioned the interaction's utility for everyday conversations when they *"struggle to find the correct words to convey what it is that I'm feeling and my emotional state means that I can't find the words to articulate myself."* (P9). Reflecting on this 'in situ' approach, another participant expressed an interest in an app version of the prototype that could be easily accessed each time they needed to sound out a problem within their day. Younger adolescents were frequently identified as a group that would particularly benefit from these types of interactions as they might struggle more with articulating their feelings and emotions. Further situations where the interaction may be helpful for young people included navigating complex social dynamics, managing mental health, and addressing challenges related to identity, puberty and work. For example, participants suggested that the chatbot could assist in managing feelings of social exclusion.

Together, these insights suggest that the specific agentic interaction cultivated for this case study, alongside the broader approachability and impartiality of LLM chatbots, may create a unique interaction environment where participants feel they can safely and effectively articulate sensitive personal experiences. In 5.2.4 we briefly explore how this articulation process was seen to benefit participants in our studies.

5.2.4 Assumption 3 - Understanding participant perceptions and perceived benefits of the interaction. Participants highlighted positive aspects related to the prototype interaction which left a strong impression on many participants (e.g., *"Overall it's a lot more positive than it is negative."* (P9).) A small minority (<5%) of participants also reported to not benefit from the interaction. One participant expressed that using the tool to articulate an experience was a *"waste of time"* while another expressed that talking to a chatbot about their experiences made them feel *"alone and unheard"*. As outlined in 5.2.1, the interaction was described as both efficient and simple/easy to use, with perceived benefits over and above the contribution of personal data (e.g., "therapeutic", "rewarding"). The articulation process outlined in 5.2.2 also demonstrates how the use of novel personas enabled empowerment, as participants felt like the chatbot

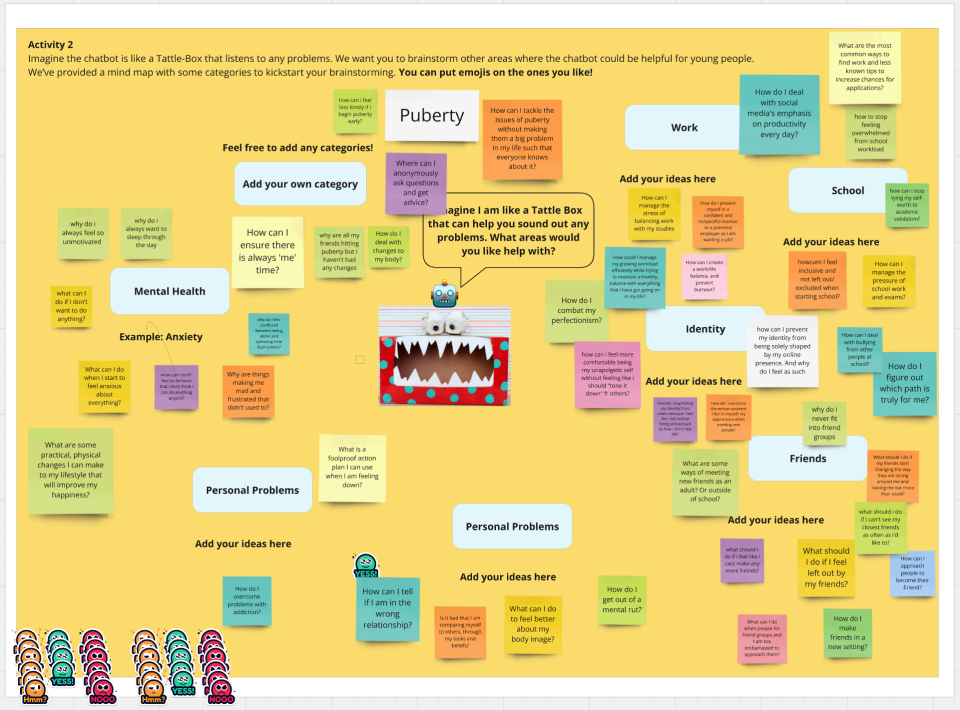

Fig. 7. Screenshot of Miro Brainstorming Activity (A4)

personas effectively captured their voices. To the extent that the overall interaction was perceived as being personally helpful, we recount several instances where participants recount having novel insights or epiphanies as a result of the interaction (see 5.2.2). To this end, we argue that the novel narrative-building approach (DA1), supported by the interaction's agentic approach (DA2), enables a unique articulation process that helps participants to make sense of and process their personal experiences, and it is this 'articulation process' which is perceived by participants as helpful and beneficial. In Study 3, we test this assumption further in a larger empirical study.

5.2.5 Overall prototype feedback. When asked about potential improvements or suggestions to the overall interaction, participants suggested that the chatbot could be made to feel more 'human', with suggestions for improving its tone to better mimic human interaction. Participants also commented on the *"repetitive"* nature of the chatbot's responding, which was at times too generic or formal. One participant also commented on the chatbot's *"limited emotional understanding"*, which they felt might hinder its ability to fully grasp the context and nuances of young people's experiences. On a related note, some participants mentioned that they would have privacy concerns about sharing deeply personal information or experiences (e.g., *"I wouldn't share deep personal things but like base issues and problems, like social media, more than happy to express that to a chatbot."*"P4). Finally, while the chatbot helped some participants consider potential solutions to their problems (*"it did make me think about the next steps"* (P10)), there was a consensus that the chatbot could benefit from including additional questions or prompts to help guide users toward solutions for their social media challenges.

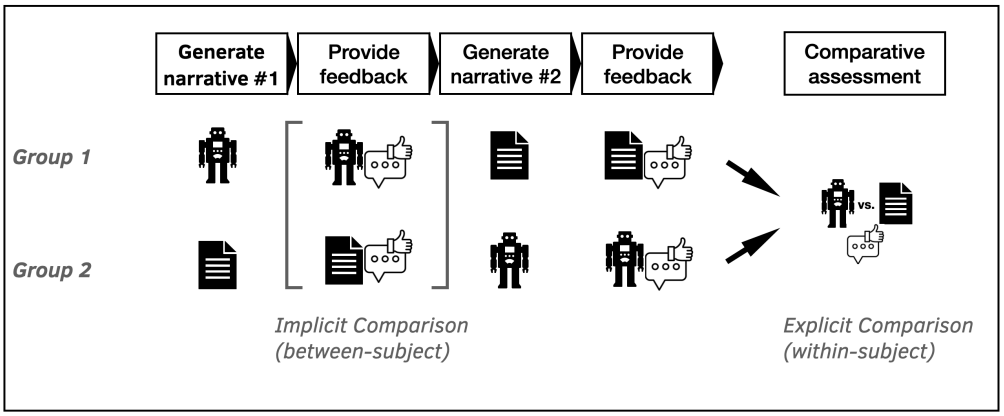

Fig. 8. Overview of empirical comparison study

6 Part 3: Comparison experiment

Based on the positive findings from studies 1 and 2, we wanted to explore if the qualitative findings that indicated user acceptance and perceived support for articulation will hold even when participants are asked to *share their actual experiences*, rather than focusing on hypothetical examples. We were also particularly interested in a more rigorous test of our third design assumption (DA3): i.e., we aimed to explore whether participants would see the articulation process as more or less simple/empowering/personally helpful when compared to the traditional open-text questionnaire question.

To explore these questions, we designed an experimental study (outlined in Figure 8) to examine how the micro-narrative elicitation flow—and the articulation interaction process it enables (outlined in 5.2.1-5.2.4)—compared to an open-form survey question (as the closest existing comparator[6]). In what follows, we focus on the user *experience* of engaging with these two methods (i.e., micro-narrative elicitation vs open-text survey), and will not report on the *content* of the stories the users produced.

6.1 Method

Participants. We recruited 269 UK-based participants on the platform Prolific. We removed 15 participants who failed at least one of two attention checks (e.g., "Place the slider between the numbers 90-100"), leaving us with a final N of 254 participants. We screened for participants between the ages of 18-20 and collected 64 participants of age eighteen, 86 participants of age nineteen, and 104 participants of age twenty ($M = 19.16$, $SD = 0.80$). We had a roughly even split of males and females ($n_{\text{male}} = 124$, $n_{\text{female}} = 122$, $n_{\text{other}} = 8$). Our sample consisted of 133 White individuals (52%), 77 Asian individuals (30%), 36 Black individuals (14%), 18 Other (7%). Participants were allowed to select multiple racial/ethnic backgrounds.

[6]We aimed to keep the open-text question as close as possible to the template used within the micro-narrative flow. Participants were informed that we were *collecting experiences about challenging situations that people have experienced on social media. Please think of a situation that you experienced on social media that you felt was challenging or troubling."* They were then asked to *"Please tell us about this experience and be sure to include details about: what happened, what was said, posted, or done, what was the context of the situation, if there is anything else that should we know about the situation, how the situation made you feel, how you reacted, and what the worst part of the situation was."*

Question Group	Question Text
Implicit comparison (between-subject)	Overall, how helpful was the [chatbot/Qualtrics form] in helping you articulate events that happened on social media?
	Overall, how difficult was it to respond to the question [posed by the chatbot/in the Qualtrics form] on the previous page?
	Overall, how valuable do you think it would be for other young people to participate in this kind of [chatbot/Qualtrics form] interaction?
	How well does the final narrative [produced by the chatbot/you entered into the Qualtrics form] accurately capture the objective circumstances of the situation (i.e., what factually happened in the situation, as opposed to your feelings about it)?
	How well does the narrative [generated by the chatbot/you entered into the Qualtrics form] accurately capture how you felt in the situation?
	If someone else who did not know anything about you or the situation read this narrative, how well would they understand your experience? [bot-generated narrative/Qualtrics form narrative]
Explicit comparison (within-subject)	Which process did you personally prefer: interacting with the bot to produce a final narrative or filling out the Qualtrics web form?
	Which process felt easier to capture what you wanted to share about your experience?
	Which process would be more appropriate when asking young people to share their stories?
	Which process was more helpful for allowing you to make sense of your experience?
	Which process did you feel better captured your voice?
	Which process was more likely to make you consider how you can address your social media challenge?
	How likely would you be to recommend each one of these experiences to a friend? [bot/Qualtrics form]
	Did you have any privacy concerns about the [bot/Qualtrics form]?

Fig. 9. Wording of the survey questions as seen by participants. Question text in square brackets indicates the text that was different in different presentations of the same question. All questions in the implicit question group were answered on a unipolar scale (i.e., a single question refers to a single input method). In the explicit comparison group, "would recommend" and "privacy concerns" were reported on unipolar scales with separate questions for each input method. The rest of the questions in the explicit comparison group were reported on a bipolar scale where one side represented that the chatbot was higher on the given dimension and the other side represented that the Qualtrics form was higher on the given dimension.

Procedure. Participants completed a three-part online study. In the first part, participants were randomly assigned to either fill out an open-text question or engage with the narrative elicitation flow described in previous sections (labelled as the 'chatbot' condition in the rest of this section). Next, they answered questions about their experience with their initially assigned condition. This allowed us to conduct between-subject analyses of the differences in people's experiences with each of the two input methods without having been biased by past experience of the other condition. We refer to this set of analyses as the "implicit comparison". In the second part, participants were assigned to do whichever activity they had not just completed (i.e., either form or chatbot) and then answered the same questions again regarding their experiences with the second condition. Finally, in the third part, participants answered a series of questions asking them to explicitly compare the two experiences on several dimensions. We will therefore refer to these dependent variables as the "explicit comparison". The wording of all questions is presented in Figure 9.

6.2 Data Analysis

6.2.1 Implicit Comparison. To analyze the responses to the implicit comparison segment of the study, we used Bayesian beta regression models estimated in the R package *brms* [4]. Beta regression models are ideal for measuring sliding scale responses, such as those employed in our study, for at least two primary reasons [62]. First, beta regression models elegantly handle skewed distributions not well-modeled by traditional linear regression. Second, unlike traditional linear models (i.e., frequentist linear regression, ANOVA, t-tests), the model-implied values of the beta regression model are bounded to the same range as the response scale. In other words, this type of model does not make predictions that are impossible given the measurement instrument. Since beta regression requires that responses fall within the

open interval of 0-1, we first scaled our 0-100 scale to the range of 0-1 and then truncated responses at the minimum or maximum of the scale by adding or subtracting .01 to the scaled response. To facilitate ease of interpretation, we present the results in the original scale that the data was collected (0-100).

In our beta regression models, we regressed the two parameters of the beta distribution, the mean parameter μ and the precision parameter ϕ, onto a binary variable indicating whether the datapoint corresponded to the Qualtrics form condition (coded as 0) or the chatbot condition (coded as 1). For the purposes of the present study, we were primarily interested in the effect of condition on μ, the mean of the beta distribution. Thus, the parameter that we report indicates the difference in means between the chatbot condition and the Qualtrics form condition, such that positive values indicate that chatbot had a greater mean and negative values indicate that the chatbot had a lower mean.

6.2.2 Explicit Comparison. In the explicit comparison portion of the study, we used two styles of response scales – unipolar scales where participants completed a separate question for each of the two input methods (chatbot and Qualtrics form) and bipolar scales where one side referred to the bot and one side referred to the Qualtrics form (with counterbalanced assignment of conditions to sides of the scale). To analyze responses on unipolar scales, we used beta regression models as described in our explanation of the data analysis for the implicit comparison questions. To analyze responses on the bipolar scales, we binarized responses for ease of interpretation and fit logistic regression models in *brms* [4]. Our logistic regression models were intercept-only models. Thus, the parameter of interest represents the log-odds that someone reported that the bot was higher on the given dimension than the Qualtrics form. We exponentiated the log-odds in our results for ease of interpretation. We used *brms* default, weakly informative priors for all models presented in the manuscript.

6.3 Results

6.3.1 Implicit Comparison Questions. We fit beta regression models to examine between-person differences between our chatbot and Qualtrics form for eliciting vignettes of negative experiences on social media. We found, generally, that the chatbot had more desirable characteristics than the Qualtrics form (see Table X). Specifically, participants found the chatbot more helpful for articulating the experience ($\mu_{\text{difference}} = 22.04$, 95% Cr.I. $[17.70, 26.27]$), less difficult to respond to ($\mu_{\text{difference}} = -5.58$, 95% Cr.I. $[-9.99, -1.17]$), more valuable for youth ($\mu_{\text{difference}} = 5.40$, 95% Cr.I. $[0.15, 10.60]$), more accurate for capturing the objective features of the situation ($\mu_{\text{difference}} = 11.75$, 95% Cr.I. $[7.43, 16.03]$), more accurate for capturing their feelings ($\mu_{\text{difference}} = 14.51$, 95% Cr.I. $[9.91, 19.07]$), and that others who read the final product would understand their experience better ($\mu_{\text{difference}} = 17.60$, 95% Cr.I. $[13.55, 21.63]$).

6.3.2 Explicit Comparison Questions. We fit beta regression models for two explicit comparison questions regarding privacy concerns and the probability of recommending each method to friends and found that people had more privacy concerns about the bot ($\mu_{\text{difference}} = 21.51$, 95% Cr.I. $[16.53, 26.56]$), but were still more likely to recommend the bot to friends ($\mu_{\text{difference}} = 23.68$, 95% Cr.I. $[17.95, 29.25]$). We fit intercept-only logistic regression models for the remaining five explicit comparison questions. Participants reported that the chatbot was easier to capture their experience (OR = 4.57, 95% Cr.I. $[3.29, 6.28]$), more appropriate for youth (OR = 2.86, 95% Cr.I. $[2.14, 3.77]$), better help in making sense of the experience (OR = 6.08, 95% Cr.I. $[4.12, 8.54]$), better captured their voice (OR = 1.79, 95% Cr.I. $[1.35, 2.30]$), and better elicited them to think about how to address their social media challenge (OR = 4.80, 95% Cr.I. $[3.35, 6.71]$).

6.3.3 Findings summary and comparison to qualitative findings. To summarise, participants preferred the chatbot tool for generating micro-narratives around their social media challenges. The chatbot was perceived as better articulating

their experiences and more accurately capturing both the objective and emotive aspects of these experiences. It was also deemed more appropriate for young people and seen to better capture participants' voices and enable sense-making of their experiences, compared to the Qualtrics form. Despite having more privacy concerns about the chatbot than the form, participants were also more likely to recommend this data collection method. Overall, these findings support the qualitative insights we drew from Studies 1 and 2 in relation to the key design assumptions. That is, participants positively experienced the LLM-support articulation process and preferred this process over traditional, non-AI interactions (DA1 and DA2). Additionally, this interaction was perceived as easier than building a narrative from scratch, better-enabling articulation in the user's voice and more helpful for sense-making and personal insights when compared to traditional data collection approaches (DA3).

7 Part 4: Informal engagements with academic and non-profit experts

By engaging with a range of established researchers and non-profit leaders across a variety of fields, our aims were two-fold: First, to informally gauge the level of interest in the idea of collecting micro-narrative datasets (assuming a reliable tool was available) and the risks they perceived in doing so; and second, to gain an initial understanding of the types of research questions/design problems—if any—that this micro-narrative collection process might be uniquely well positioned to address. Given this approach, the rest of this section is not to be read as a traditional interview study (which would require a paper of its own), but rather as a set of informal engagements that served to help us, as designers, to start mapping out the potential use-cases and associated risks, which could be explored in future work. We note that these engagements took place *in parallel* to the studies described in Sections 4-6 – i.e., , the expert's reactions reported below included their own experiences with the prototype, prior to us having the positive data from the youth case study to report.

We spoke with 14 established researchers (median citations 13.1k; range 1850-79500) across a range of international institutions[7], and a spectrum of research domains[8] that could benefit from such a tool. We also engaged with CEO/VP level staff from four non-profits across UK, US, and Mexico, each serving more than 20k users, and several counting their reach in millions of young people.

Procedure. Each expert was provided with a short email explanation and the opportunity to try out the current prototype over email, followed by at least 30min long zoom call (or in-person conversation). Surprisingly, seven of the researchers and all of the four non-profits were immediately interested in exploring how the tool could be embedded into their ongoing research / data collection following the call. In these cases, the initial conversations turned into a thread of follow-up calls and emails (which are still ongoing at the point of submission). During the initial call, we were interested in understanding the experts' immediate reactions to their interactions with the bot, how they imagined the collected data would be similar or different to interviews / surveys (or other methodologies they use on a regular basis), as well as what were their perceived benefits *and* risks if this approach was used in their contexts.

Observations and insights. Overall, the experts seemed to value the ability to collect personal stories at scale as well as—often—showing a sense of surprise in how well the AI was able to combine their own words into fluent narratives. These insights then nearly always naturally progressed to reflections on how such data could be useful in their work. In

[7]Cambridge, Harvard, King's College London, LSE, University of Melbourne, University of Michigan, Monash University, Northwestern University, University of Nottingham, University of Oxford, University of British Columbia, University of California Irvine, University of Washington
[8]The broad domains included psychology, clinical psychology (including researchers working on self-harm/suicide interventions), implementation science, health behaviour change, and HCI.

what follows, we structure around four broad 'types' of use-cases that consistently emerged across our various expert conversations.

(1) Identifying **'canonical' stories**: The first type aligned with our original challenge of capturing personal narratives that address a specific, well-defined question from a large population. For example, some of our experts mentioned needing to understand the most salient challenge with accessing help (or lack thereof) for young people experiencing self-harm; or capturing examples of emergent good practice around large-scale behaviour change intervention implementation. In such examples, the micro-narratives were to be used to help distil the most commonly shared experiences, which would then be taken forward to guide further research (e.g., intervention development).

(2) Gathering **multi-story collections**: Some of the experts were interested in the opportunity to collect multiple stories from each participant over a longer period (e.g., daily or weekly), as a way to create a more holistic understanding of participant experiences. For example, one of our experts was interested in collecting a series of micro-narratives to understand more about what it feels like to live with obesity for a specific marginalised US population – including stories about 'a time they felt like giving up', 'a time when someone said or did something that really helped on their journey', or a 'time when they felt pressured to lose more weight then felt healthy for them'. We note that, ideally, at least some such target story 'stems' within the collection would be co-produced by members of the community to capture the type of narratives that the participants themselves want to share.

(3) Micro-narrative as **part of interventions**: In these cases, our experts were interested in collecting individual stories directly as part of the active intervention components, rather than 'just' as an empirical data collection. They envisioned such micro-narratives as a novel approach to amplify existing intervention approaches (e.g., as an input into a weekly patient-therapist conversation, or driving a goal setting process in behavioural change), or serve as an intervention of its own (e.g., as part of driving reflective practice or as a cognitive distancing intervention).

(4) Capturing **intervention outcomes**: Finally, some experts perceived micro-narratives as an innovative approach to study the effects of interventions at scale (e.g., process analysis capturing most people within an intervention), while remaining highly personalised and open-ended to capture patients' needs; or even driving intervention adaptation (e.g., as an input into JITAI decision algorithms).

The excitement about the *potential* of similar micro-narrative techniques was understandably tempered by safe-guarding and data protection questions, as well as concerns about potential hesitations from respective IRB / ethics committees – at least until this LLM-driven method is "standardised". Our perception was that the experts saw these issues as crucial steps that will require deep and thoughtful work as this methods will continue to be developed; but didn't seem to see any risks that would led to insurmountable concerns. We discuss the expert perspectives and our own reflections from the empirical studies on risk and ethical questions in more detail in the Future Research Agenda subsection of the discussion (Section 8).

8 Discussion

The purpose of this paper has been to try to develop a novel data collection technique— micro-narrative elicitation—that would provide a new way of approaching the trade-off between the richness of the collected datasets, and the required effort from researchers / participants to do so. In what follows, we first briefly revisit our empirical findings and then move on to discussing the (many!) remaining open questions that will need to be resolved.

Summary of results across studies. Our design explored the potential of a novel human-AI flow to scaffold users' articulation of short narratives about their lived experience. Three key design assumptions drove our development and design work: First, the *users' articulation process* can be based on a template-based vignette structure – enabling the researchers to specify the structure of the story, but allowing the participants to fill it in with their own words (DA1). Second, that the *cognitively burdensome parts of this process could be supported through LLM-powered components*, while fully *retaining the participants' words and control* over the resulting narrative (DA2). And finally, that *such a process would be ideally perceived as simple, empowering, and insightful* by the participants (DA3).

Across our three mixed-method studies (an initial pilot (N=100), 2-week co-design (N=30), and an experiment (N = 254)) the findings demonstrated the acceptability & potential for micro-narrative elicitation as a new approach to collecting narrative data about participant lived experiences. Overall, the participants' feedback suggest two plausible overarching benefits of this method: First, it seems to represent an easy and efficient way for participants to articulate a personal experience in a way that 'makes sense', and is also often perceived as inherently enjoyable (DA1, DA2). Second, this process, of articulating an experience through a short agentic interaction, appears to help participants to make sense of or emotionally reflect upon the experience – indicating the potential to generate additional benefits over and above the contribution of their personal data (DA2,DA3).

Moreover, our informal engagements with experts in non-profit and academic domains suggested the potential for such a narrative elicitation interactions could have potential across multiple research domains. In particular, the range of use-cases that our experts were interested to explore was unexpected, and for many of our experts included the wish to explore the use of such narrative methods as part of interventions (rather than as a data collection method only). Finally, we note that our initial attempts to adapt the tool for alternative narratives were surprisingly simple – only requiring the change of the list of questions in Data Collection component, and narrative template in the Narrative Creation LLM. This could be accomplished in under 1 hour through manual changes, and we are currently working on refactoring the codebase to enable full automation of the process.

Relationship to other data collection methods. We want to be clear that we *do not* expect the micro-narrative elicitation method proposed here to replace any existing approach to data gathering; or attempt to make comparisons in terms of their value for design in general. Instead, this approach should be seen just as an *additional tool*, one that is complementary to well-known methods, and that might—at times—be well-suited to particular design questions.

In our view, any potential benefits of this method stem simply from a different approach to balancing data richness vs. effort: on one hand, the proposed process is highly constrained by its ability to *only* capture a highly pre-specified aspect of the user's experiences (that fit the proposed vignette template); on the other, it is exactly this narrow focus that then allows us to provide streamlined support to simplify users' articulation of their stories. Such a scaffolded approach to narrative articulation then raises a number of open questions about the ethics and epistemological considerations around the nature of the data collected, which is what we turn to in the next section.

Open questions and research agenda

Safeguarding and ethical use: Any data collection method that aims to capture participants' experiences around sensitive issues must carefully consider key questions around data safeguarding and other ethical risks – and many well-researched approaches to safeguarding protocols are already available within the psychological literature (e.g., in domains such as self-harm or suicidal ideation, cf., [33]). Similarly, substantive literature has already engaged with the questions of the use of AI in data collection (cf., [47, 59]) as well as other highly sensitive contexts (such as AI use in

mental health therapy [53]). Based on our discussions with experts and our own reflection, we would like to highlight three main directions that any use of this—or similar methods—should consider.

Firstly, when probing sensitive personal experiences, particularly those surrounding mental well-being, there is a risk of participants disclosing sensitive information. The magnitude of such risk clearly differs across topics (e.g., asking for teachers' best practices is less risky than collecting stories of mental health help-seeking). However, such risk will increase with larger sample sizes and is particularly amplified for vulnerable populations, such as youth and adolescents (cf., for example [48]). Researchers should therefore implement appropriate safeguarding protocols alongside the deployment of these tools to mitigate disclosure risks and ensure researchers respond appropriately and efficiently to disclosures. This might involve automatic monitoring[9] (cf., [26]) or potentially even including a trained therapist 'on-call' to review the data as it is being collected, which is common in highly sensitive clinical settings including large-scale out-patient trials [42].

A second open question concerns LLM responses and the potential for LLMs to respond insensitively or detrimentally to participants' experiences. While our approach inherently mitigates this risk by strongly scoping the interaction (to asking specific, pre-determined questions), there is again a need to consider additional safeguarding, particularly for vulnerable and marginalised populations. This is an active area of research [30, 66], with a wide range of techniques potentially available depending on the research contexts and specific risks.

Finally, the ethics of developing a 'listener' AI have not been fully explored in research and further exploration is needed to understand how these tools may engage and affect the participants who use them – and in the context of this work, also shape the narratives being produced (cf., [18]. This is particularly in light of the present findings, where some youth described the interaction as therapeutic/supportive, despite this not being a design aim. We see this as a crucial area for future research, and one which can draw on existing work around therapeutic chatbots – cf., [13] for an extensive recent review.

Epistemological questions: The conversations with experts opened a range of questions around how one could understand, interpret, and analyse the micro-narrative data. On one hand, more research is needed to understand the extent to which the data—co-created as a part of human-AI workflow—is indeed capturing what the participant has *meant to* articulate; and how we should interpret such data in situations where they would not have been able to describe such a story without such help. Further, there are epistemological challenges of working out how one might analyse/interpret large datasets (e.g., 100s of micro-narratives) and how this relates—if at all—to the more traditional iterative and discursive approaches used in HCI (cf., Section 4). Finally, the resulting narrative is only part of the data that could be captured. It is possible for the system to also collect the highly structured 'provenance' that led to the narrative[10] – all of which could be used for computational analysis, for example similarly to what is currently done on large-scale Reddit datasets.

Broader design questions. In principle, an abstracted description of the design work here is focused on a specific cognitive flow (i.e., articulating own experience into a narrative form that includes predetermined aspects of the situation), which has been seen as important but also hard for many participants to do. The design itself then 'only' reframed such cognitive flow interaction into components that seemed 'easy' for the human to do (answer questions about fragments; make a choice between scenarios as to which one seems closest to own perception; ask for adaptation)

[9]For example, implemented as one (or more) parallel modules sitting on top of Data Collection LLM component, checking the dialogue for risks, and invoking safeguarding procedures—such as diverting to another interaction flow, providing help resources, or alerting a human therapist—if necessary.
[10]Including, for example, the verbatim text written by the participants, the extracted content LLM used, the choices made by the participants, and any follow-up adaptation requests / edits

while ensuring that the remaining components were 'easy' to do for the computational system (ask X questions in a row; extract information; combine information following a template). Some of the envisioned use-cases by the experts are already pointing to extensions that could rely on similar design patterns. For example, the suggestions to use the micro-narratives as a cognitive distancing intervention (cf., [10]) would likely include retaining the story generation (current process), but then extending it with components that could, e.g., address the challenging parts of the 'decentering' metacognitive process; as a core-but-difficult cognitive flow within the existing intervention approach.

We expect that there must be many other important cognitive flows that could be decomposed in analogous ways, and lead to innovative HCAI systems that are based on a deep understanding of the learning / cognitive challenges that users experience (cf., [52]) and the understanding of the strengths that (agentic) LLM systems could bring into the process.

9 Conclusion

This paper introduces a new data collection method—micro-narratives—which could help collect rich but narrowly scoped qualitative data at scale. Specifically, the method aims to empower participants to easily articulate their stories in their own voice, regardless of their writing ability. To accomplish this, the method leverages AI chaining to enable the design of a scaffold that breaks the creation of such narratives into steps that are easy for individuals to accomplish, automates aspects of narrative creation that are difficult, while giving individuals agency to revise the resulting narratives to make them their own. Both qualitative and quantitative results showcase the acceptability and feasibility of this data collection approach within a particular case study; while expert engagements point to a wide range of potential additional applications.

References

[1] Noorhan Abbas, Thomas Pickard, Eric Atwell, and Aisha Walker. 2021. University student surveys using chatbots: artificial intelligence conversational agents. In *International Conference on Human-Computer Interaction*. Springer, 155–169.

[2] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173951

[3] Kirsten Boehner, Janet Vertesi, Phoebe Sengers, and Paul Dourish. 2007. How HCI interprets the probes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1077–1086. https://doi.org/10.1145/1240624.1240789

[4] Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80 (2017), 1–28.

[5] Susanne Bødker. 2015. Third-wave HCI, 10 years later—participation and sharing. *interactions* 22, 5 (Aug. 2015), 24–31. https://doi.org/10.1145/2804405

[6] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[7] Thomas Anthony Chun Hun Chan, Jason Man-Bo Ho, and Michael Tom. 2023. Miro: Promoting collaboration through online whiteboard interaction. *RELC Journal* (2023), 00336882231165061.

[8] Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods* 47, 4 (2015), 1178–1198.

[9] A Crabtree, T Hemmings, T Rodden, K Cheverst, K Clarke, G Dewsbury, J Hughes, and M Rouncefield. 2004. Designing with care: Adapting cultural probes to inform design in sensitive settings. In *Proceedings of the 2004 Australasian Conference on Computer-Human Interaction (OZCHI '04)*. Ergonomics Society of Australia, Brisbane, Australia, 4–13.

[10] Quentin Dercon, Sara Z. Mehrhof, Timothy R. Sandhu, Caitlin Hitchcock, Rebecca P. Lawson, Diego A. Pizzagalli, Tim Dalgleish, and Camilla L. Nord. 2024. A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychological Medicine* 54, 2 (2024), 327–337. https://doi.org/10.1017/S0033291723001587

[11] Giulia Di Fede, Davide Rocchesso, Steven P Dow, and Salvatore Andolina. 2022. The idea machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*. 623–627.

[12] Spencer C Evans, Michael C Roberts, Jared W Keeley, Jennifer B Blossom, Christina M Amaro, Andrea M Garcia, Cathleen Odar Stough, Kimberly S Canter, Rebeca Robles, and Geoffrey M Reed. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application

in ICD-11 field studies. *International journal of clinical and health psychology* 15, 2 (2015), 160–170.

[13] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024. The Ethics of Advanced AI Assistants. arXiv:2404.16244 [cs.CY] https://arxiv.org/abs/2404.16244

[14] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural Probes. *interactions* 6, 1 (Jan. 1999), 21–29. https://doi.org/10.1145/291224.291235

[15] William W. Gaver, Andrew Boucher, Sarah Pennington, and Brendan Walker. 2004. Cultural Probes and the Value of Uncertainty. *Interactions* 11, 5 (Sept. 2004), 53–56. https://doi.org/10.1145/1015530.1015555

[16] Carol Haigh and Pip Hardy. 2011. Tell me a story—a conceptual exploration of storytelling in healthcare education. *Nurse education today* 31, 4 (2011), 408–411.

[17] Zoë Hobson, Julia A Yesberg, Ben Bradford, and Jonathan Jackson. 2021. Artificial fairness? Trust in algorithmic police decision-making. *Journal of experimental criminology* (2021), 1–25.

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

[19] Juan Pablo Hourcade. 2015. *Child-Computer Interaction* (eds ed.). CreateSpace Independent Publishing Platform; First Edition edition, online.

[20] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Ft. Lauderdale Florida USA, 17–24. https://doi.org/10.1145/642611.642616

[21] Seray B Ibrahim, Alissa N. Antle, Julie A. Kientz, Graham Pullin, and Petr Slovak. 2024. A Systematic Review of the Probes Method in Research with Children and Families. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference* (Delft, Netherlands) *(IDC '24)*. Association for Computing Machinery, New York, NY, USA, 157–172. https://doi.org/10.1145/3628516.3655814

[22] Mizuko Ito. 2013. *Hanging out, messing around, and geeking out: Kids living and learning with new media.* The MIT press.

[23] Betul Keles, Annmarie Grealish, and Mary Leamy. 2024. The beauty and the beast of social media: an interpretative phenomenological analysis of the impact of adolescents' social media experiences on their mental health during the Covid-19 pandemic. *Current Psychology* 43, 1 (2024), 96–112.

[24] Christoph Kern, Frederic Gerdon, Ruben L Bach, Florian Keusch, and Frauke Kreuter. 2022. Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making. *Patterns* 3, 10 (2022).

[25] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[26] Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-Woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2024. MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients' Journaling. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), 1–20. https://doi.org/10.1145/3613904.3642937 arXiv:2310.05231

[27] Philippe Laban, Jesse Vig, Marti A Hearst, Caiming Xiong, and Chien-Sheng Wu. 2023. Beyond the chat: Executable and verifiable text-editing with llms. *arXiv preprint arXiv:2309.15337* (2023).

[28] H Lahti, M Kulmala, N Lyyra, V Mietola, and L Paakkari. 2024. Problematic situations related to social media use and competencies to prevent them: results of a Delphi study. *Scientific Reports* 14, 1 (2024), 5275.

[29] Jean Lave and Etienne Wenger. 1991. *Situated learning: legitimate peripheral participation.* Cambridge University Press, Cambridge [England] ; New York.

[30] Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health* 11, 1 (2024), e59479.

[31] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), 1–25. https://doi.org/10.1145/3613904.3642625

[32] Sonia Livingstone and Alicia Blum-Ross. 2020. *Parenting for a digital future: How hopes and fears about technology shape children's lives.* Oxford University Press, USA.

[33] Elizabeth E Lloyd-Richardson, Stephen P Lewis, Janis L Whitlock, Karen Rodham, and Heather T Schatten. 2015. Research with adolescents who engage in non-suicidal self-injury: ethical considerations and challenges. *Child and adolescent psychiatry and mental health* 9 (2015), 1–14.

[34] Haley MacLeod, Ben Jelen, Annu Prabhakar, Lora Oehlberg, Katie A Siek, and Kay Connelly. 2016. Asynchronous remote communities (ARC) for researching distributed populations.. In *PervasiveHealth*. 1–8.

[35] Haley MacLeod, Ben Jelen, Annu Prabhakar, Lora Oehlberg, Katie A Siek, and Kay Connelly. 2016. Lessons learned from conducting group-based research on facebook. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 804–815.

[36] Bahar Mahmud, Guan Hong, and Bernard Fong. 2023. A Study of Human–AI Symbiosis for Creative Work: Recent Developments and Future Directions in Deep Learning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 2, Article 47 (sep 2023), 21 pages. https://doi.org/10.1145/3542698

[37] Tuuli Mattelmäki and Katja Battarbee. 2002. Empathy Probes. In *PDC 02 Proceedings of the Participatory Design Conference*. CPSR, Malmo, Sweden, 266–271. https://ojs.ruc.dk/index.php/pdc/article/view/265

[38] Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence* 6 (2024), 100199.

[39] Angie L Miller and Amber D Dumford. 2014. Open-Ended Survey Questions: Item Nonresponse Nightmare or Qualitative Data Dream? *Survey Practice* 7, 5 (2014), 1–11. https://doi.org/10.29115/sp-2014-0024

[40] Maria Poulou. 2001. The role of vignettes in the research of emotional and behavioural difficulties. *Emotional and behavioural difficulties* 6, 1 (2001), 50–62.

[41] Stanislav Pozdniakov, Jonathan Brazil, Solmaz Abdi, Aneesha Bakharia, Shazia Sadiq, Dragan Gasevic, Paul Denny, and Hassan Khosravi. 2024. Large Language Models Meet User Interfaces: The Case of Provisioning Feedback. *arXiv preprint arXiv:2404.11072* (2024).

[42] Amie Randhawa, Grace Wood, Maria Michail, Miranda Pallan, Paul Patterson, and Victoria Goodyear. 2024. Safeguarding in adolescent mental health research: navigating dilemmas and developing procedures. *BMJ open* 14, 2 (2024), e076700.

[43] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of humanization to survey chatbots: Change in chatbot perception, interaction experience, and survey data quality. *Computers in Human Behavior* 126 (2022), 107034.

[44] Yvonne Rogers and Paul Marshall. 2017. Research in the Wild. *Synthesis Lectures on Human-Centered Informatics* 10, 3 (April 2017), i–97. https://doi.org/10.2200/S00764ED1V01Y201703HCI037

[45] Helen Sampson and Idar Alfred Johannessen. 2020. Turning on the tap: the benefits of using 'real-life' vignettes in qualitative research interviews. *Qualitative Research* 20, 1 (2020), 56–72. https://doi.org/10.1177/1468794118816618

[46] Viktor Schønning, Gunnhild Johnsen Hjetland, Leif Edvard Aarø, and Jens Christoffer Skogen. 2020. Social media use and mental health and well-being among adolescents–a scoping review. *Frontiers in psychology* 11 (2020), 1949.

[47] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2023. ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events. *arXiv* (2023). https://doi.org/10.1145/3613904.3642152 arXiv:2309.12244

[48] Siobhan Sharkey, Ray Jones, Janet Smithson, Elaine Hewis, Tobit Emmens, Tamsin Ford, and Christabel Owens. 2011. Ethical practice in internet research involving vulnerable people: lessons from a self-harm discussion forum study (SharpTalk). *Journal of medical ethics* 37, 12 (2011), 752–758.

[49] Jessica Sheringham, Isla Kuhn, and Jenni Burt. 2021. The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. *BMC Medical Research Methodology* 21, 1 (2021), 81. https://doi.org/10.1186/s12874-021-01247-4

[50] Joongi Shin, Michael A. Hedderich, Bartłomiej Jakub Rey, Andrés Lucero, and Antti Oulasvirta. 2024. Understanding Human-AI Workflows for Generating Personas. *Designing Interactive Systems Conference* (2024), 757–781. https://doi.org/10.1145/3643834.3660729

[51] Jesper Simonsen and Toni Robertson (Eds.). 2013. *Routledge international handbook of participatory design*. Routledge, London. OCLC: 818827037.

[52] Petr Slovak and Sean A. Munson. 2024. HCI Contributions in Mental Health: A Modular Framework to Guide Psychosocial Intervention Design. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 692, 21 pages. https://doi.org/10.1145/3613904.3642624

[53] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 5 (2020), 1–53.

[54] Jukka Törrönen. 2002. Semiotic theory on qualitative interviewing using stimulus texts. *Qualitative research* 2, 3 (2002), 343–362.

[55] Dominique Tremblay, Annie Turcotte, Nassera Touati, Thomas G Poder, Kelley Kilpatrick, Karine Bilodeau, Mathieu Roy, Patrick O Richard, Sylvie Lessard, and Émilie Giordano. 2022. Development and use of research vignettes to collect qualitative data from healthcare professionals: A scoping review. *BMJ open* 12, 1 (2022), e057095.

[56] Dominique Tremblay, Annie Turcotte, Nassera Touati, Thomas G Poder, Kelley Kilpatrick, Karine Bilodeau, Mathieu Roy, Patrick O Richard, Sylvie Lessard, and Émilie Giordano. 2022. Development and use of research vignettes to collect qualitative data from healthcare professionals: a scoping review. *BMJ Open* 12, 1 (2022), e057095. https://doi.org/10.1136/bmjopen-2021-057095

[57] Patti M Valkenburg. 2022. Social media use and well-being: What we know and what we need to know. *Current opinion in psychology* 45 (2022), 101294.

[58] Amber van der Wal, Patti M Valkenburg, and Irene I van Driel. 2024. In Their Own Words: How Adolescents Use Social Media and How It Affects Them. *Social Media+ Society* 10, 2 (2024), 20563051241248591.

[59] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–35. https://doi.org/10.1145/3637364 arXiv:2301.05843

[60] Emily Weinstein and Carrie James. 2022. *Behind their screens: What teens are facing (and adults are missing)*. MIT Press.

[61] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[62] Jerzy Wieczorek and Sam Hawala. 2011. A bayesian zero-one inflated beta model for estimating poverty in us counties. In *Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association*. 2812–2815.

[63] Peter Wright and John McCarthy. 2010. *Experience-centered design: designers, users, and communities in dialogue*. Number 9 in Synthesis lectures on human-centered informatics. Morgan & Claypool Publ, San Rafael, Calif. OCLC: 700336767.

[64] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *CHI Conference on Human Factors in Computing Systems* (2022), 1–22. https://doi.org/10.1145/3491102.3517582 arXiv:2110.01691

[65] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37. https://doi.org/10.1145/3381804 arXiv:1905.10700

[66] H Yu and Stephen McGuinness. 2024. An experimental study of integrating fine-tuned LLMs and prompts for enhancing mental health support chatbot system. *Journal of Medical Artificial Intelligence* (2024), 1–16.

[67] J.D. Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (2023), 2206–2220. https://doi.org/10.1145/3563657.3596138

[68] Sena Çerçi, Marta E. Cecchinato, and John Vines. 2021. How Design Researchers Interpret Probes: Understanding the Critical Intentions of a Designerly Approach to Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3411764.3445328