The Promise and Peril of On-Device AI for Conservation Work

Cynthia Dong
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, Washington, USA
cyndo@cs.washington.edu

Emmanuel Azuh Mensah
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, Washington, USA
emazuh@cs.washington.edu

Vaishnavi Ranganathan
Microsoft Research
Seattle, Washington, USA
vaishnavi.ranganathan@microsoft.com

Kurtis Heimerl
Paul G. Allen School of Computer Science and Engineering
University of Washington
Seattle, Washington, USA
kheimerl@cs.washington.edu

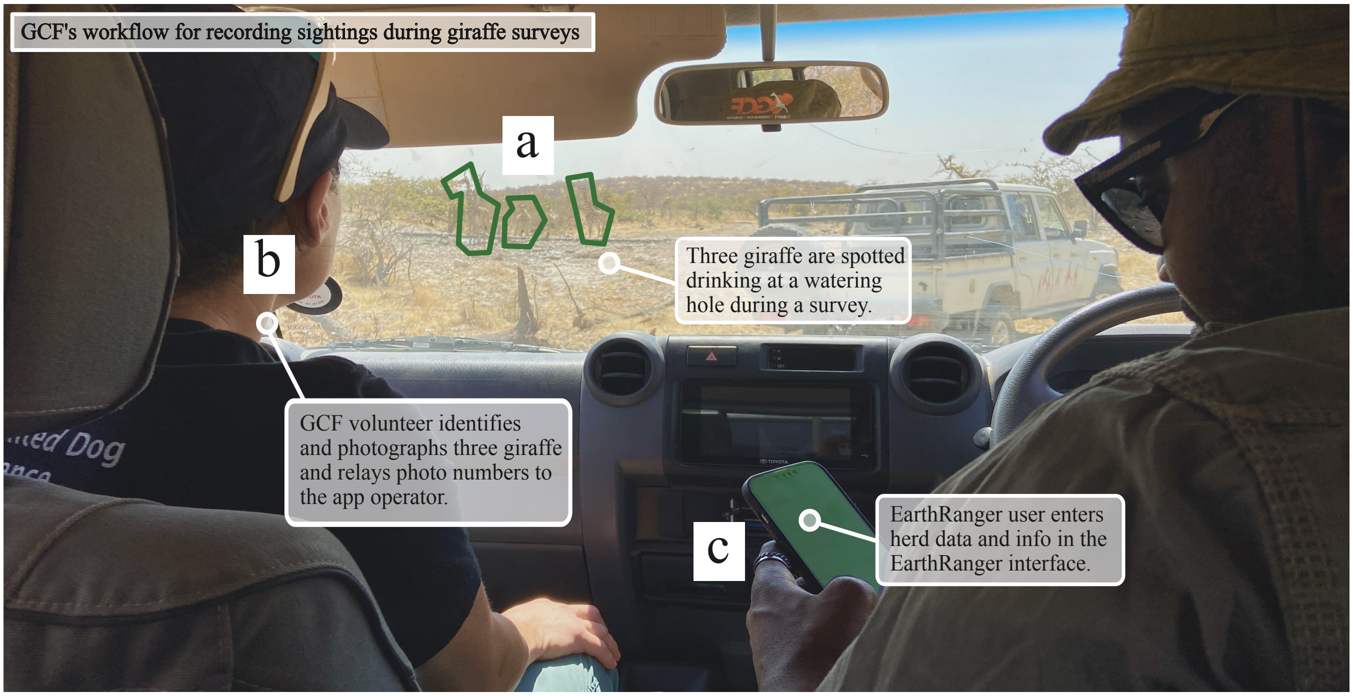


Figure 1: GCF field conservationists and volunteers surveying a giraffe herd at a watering hole. (a) Three giraffe, the target of the survey. (b) Volunteer photographing the three giraffe and calling out photo numbers to the GCF field conservationist. (c) GCF field conservationist entering data into an EarthRanger event form.

Abstract

At the heart of conservation are the field staff who study and monitor ecosystems in challenging environments. Recent advances in AI models raise the question of whether LLM assistants could improve the experience of data collection for these staff. However, on-device AI deployment for conservation field work poses significant challenges, and is understudied. To address this gap, we conducted semi-structured interviews, surveys, and participant observation with partner conservancies in the Pacific Northwest and Namibia to better understand the field work context. We employ speculative methods through the lens of technology acceptance theory to critically analyze how on-device AI would affect field work, by developing an on-device transcription-language model pipeline, which we built atop of EarthRanger, a widely-used, open-source conservation platform. Our findings suggest that although on-device LLMs hold some promise for field work, the infrastructure required by current on-device models clashes with the reality of resource-limited conservation settings.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791359

CCS Concepts

• **Human-centered computing → Mobile phones**; *Collaborative and social computing systems and tools*; • **Applied computing →** Computers in other domains.

Keywords

Wildlife Conservation, Rural Computing, AI-Assisted Data Entry, On-Device AI, ICT4D

ACM Reference Format:

Cynthia Dong, Emmanuel Azuh Mensah, Vaishnavi Ranganathan, and Kurtis Heimerl. 2026. The Promise and Peril of On-Device AI for Conservation Work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3772318.3791359

1 Introduction

The overwhelming consensus amongst researchers is that the planet faces significant biodiversity loss: anywhere from 20% to 50% of species may go extinct by the year 2100, with significant ramifications for the health of people and the environment [24]. However, experts also agree — through funding conservation projects, we can significantly reduce biodiversity loss over the coming decades [50].

The advancement of tracking and sensor technologies has progressed alongside the adoption of computer-based advances in human-generated wildlife observations, leading to an explosion in the volume and diversity of data available, especially for big animal tracking [25, 51]. In parallel, conservancies have expedited their ability to gather and understand data from observations of flora and fauna across broad geographic areas through technology.

Taken together, these advances have broadened and accelerated the flow of data from the field. While this new flood of information may alleviate conservation's data scarcity problem, it also generates new challenges: namely, those of large-scale unstructured data management and analysis [25]. Many conservation organizations lack the personnel capacity to manage the data generated, especially given widespread staffing shortages [2, 43]. Field staff also report "excessive safety and health risks" on-the-job, with the World Wildlife Fund finding that 30% of field staff received inadequate training and preparation for their job in 2019 [31]. Through informal conversations with several organizations, we learned that many field staff work with and/or are drawn from local communities, and may lack digital and textual literacy, complicating data-gathering. These challenges compound burdens upon overtaxed staff, especially in data collection, which requires naturalist expertise and is thus difficult, if not impossible, to scale.

In recent years, proposals to use AI systems (especially digital assistants based on Large Language Models (LLMs)) to increase workforce efficiency have sprung up across many sectors — corporate meetings, healthcare, and even conservation [4, 14, 19, 46]. As users in these settings generally have access to abundant electricity, stable internet, and top-of-the-line devices, developers commonly make assumptions about resources available. However, given the uneven and often limited computing infrastructure available in the remote areas where field staff work, existing research on digital assistants, which often rely on API calls to cloud-hosted services, may no longer hold, indicating a need for edge inference. There are other benefits as well: on-device inference can limit the leakage of sensitive identifying location and reduce conflicts with government data policies. Conservation organizations operate across national borders — for example, some of our partners work across multiple countries, from Angola to Namobia to South Africa. On-device processing keeps data in-country, partially circumventing the complicated data governance and privacy questions inherent to working with protected species across political landscapes.

In this work, we seek to evaluate the potential of AI systems as digital assistants in the low-resource setting of conservation work. We specifically explore the potential of *on-device* LLMs to improve data collection for field staff, allowing for easier digitization of field observations collected using their specialized ecological knowledge. To do so, we worked with a widely-used, conservation-focused data management platform provider, EarthRanger [51] (ER), and a few of their partner organizations, notably Panthera and the Giraffe Conservation Foundation (GCF).

As we began work on this project, we found that organizations using ER vary widely. This diversity of potential partners meant that our levels of intuition for user experiences similarly varied widely and needed more grounding. At the same time, we were operating within the underexplored domain of on-device LLM *deployments*, which required us to consider not only the usability of any proposed system but also the necessary supporting infrastructure. As such, this work employs a broad array of methodologies, which allowed us to holistically explore the varied potential of on-device AI for conservation. Travel to and coordination with conservation partners can be very tricky due to distance, seasonality, and remoteness; combining different methods also helped compensate for these challenges by providing alternate methods of understanding.

To provide ourselves with the contextual information necessary to reflect on insights gained from our technical validation, we conducted semi-structured interviews, a survey, and participant observation sessions with field staff in the Washington Olympic Peninsula and the Etosha Heights Game Reserve (EHGR), using the ER platform as a case study to answer the question: **RQ1: What elements are critical to developing context-suitable and maintainable conservation technologies for data work?**

In parallel, curious about the tradeoffs one would have to make between model size and performance, we explored the feasibility of on-device LLMs through the development of an on-device speech-to-text plus language model pipeline which allows field staff to dictate field observations, which are converted to ER Event forms. Not only did this validation allow us to test the technical capabilities of such a system given the current state of LLM technology, the process combined with analysis through a technology acceptance lens allowed us to probe potential benefits and challenges for the growing movement towards the deployment of LLMs on-device even outside of conservation, as small, fine-tuned LLMs become a potential model for agentic workflows [3].

Beyond traditional qualitative methods, we applied speculative and technology acceptance methods to envision how such a technology might interact with infrastructure that does not yet exist [16, 49, 53, 54]. Applying these frameworks allowed us to think *beyond* research prototype evaluation and into questions around infrastructure and longevity.

Using our system to probe the practical challenges of deploying on-device LLMs in-the-field, we sought to answer the following research question through our aforementioned qualitative methods and a meta-analysis of our technical validation. **RQ2: For conservation field staff, where do LLMs hold promise, and where do LLMs struggle to align with existing best practices in developing field-deployed conservation technology?**

From these research questions, we present a novel understanding of how new modes of digital data collection function in the context of conservation field work. We also contribute analysis of how on-device AI deployment within this context could, at its best, improve the experience of data collection for conservation — but at its worst, could upset the systems which conservation technology providers have developed to ensure that their offerings remain suitable and maintainable at low costs to the user.

2 Related Work

We provide a review in four parts: the first, focusing on existing conservation technologies and how they attempt to unify data collection while reducing labor burden, as well as what researchers see as significant challenges. The second part of our review looks at technically similar AI systems, identifying areas of success but also establishing how these systems may fail for conservation field work. We then describe user acceptance and speculative methods, which we apply as analysis and design methodologies.

2.1 Advances in Conservation Technology

Over the past two decades, the technology and tools used for conservation have expanded, from genetic barcoding to remote sensing [41]. Our interest lies in understanding how the use of technology by field staff currently affects in-the-field data entry. Hence, we focus on data collection and management tools.

There are several large platforms which dominate the conservation space: one is from our partner organization, EarthRanger (ER) [51], which aggregates and analyzes disparate real-time data sources (including data entry forms), while also visualizing data geographically. CyberTracker [28], which integrates with ER, focuses on tracking and visual functionality for oralate users, especially those using indigenous tracking methods. Other data collection platforms include citizen science platforms such as iNaturalist [22], Wildlife Insights [1] and eBird [38], which allow non-expert users to capture field recordings and attach text-based observations, with some downstream cloud AI analysis available. On the data management side, Movebank [25] provides database software towards the integration of animal movement data. And though not as widely-deployed as other systems, Muashekele et al. propose a mobile application that enables in-the-field data entry using audio recordings, which was developed through a codesign process with guards at protected wildlife areas across Namibia [33].

More broadly, several reviews of advances in conservation technology have identified trends and concerns around technology suitability and maintainability. In 2015, Pimm et al. identified many emerging conservation technologies which have the potential to reduce the labor burden on field staff and data analysts alike [41], but also established concerns: power imbalance between biodiverse conservation regions and wealthy, tech-producing countries; the high cost of new technologies; and the importance of recognizing that technology is but a tool, not a full solution for many problems [41]. In 2021, Lahoz-Monfort et al. provided a review of more recent advances, identifying further shifts away from more labor-intensive technologies and towards automation [27]. They also identify two important requirements for future contributions to conservation technology: a need for technologies which function given the harsh real-world constraints of the conservation environment, and the careful design of new systems to avoid falling into colonial patterns of dependence on software-producing nations [27]. Brammer et al. came to a similar conclusion in their work studying the use of digital data entry in participatory wildlife monitoring, cautioning that digitization may worsen data control and privacy, and increase conservationists' reliance on external technical support [5].

In the context of data collection in conservation, these works illustrate a general movement towards greater efficiency and digitization of field work, while also highlighting structural pitfalls which arise with technological advances. The integration of AI into these systems could reduce the labor of field work, but could also exacerbate structural problems. We build on these observations to explore the previously unstudied effects of integrating AI.

2.2 AI Assistants for Data Labor

While AI assistants have not yet been widely researched in conservation, they have been well-studied in the domain of healthcare, which shares similar staffing problems [26, 39]. There exists a fundamental tension in healthcare between the high-skill tasks clinicians have practiced for (diagnostics, for example) and the paperwork they must complete, largely for documentation and billing [45]. Many clinicians find that the latter encroaches upon the former in both time and focus. "AI scribes" have been proposed to alleviate the burden of paperwork on the healthcare system [11, 14, 32, 46], which, like our system, use both audio transcription and LLMs. Although the healthcare context diverges from ours, we find that reviewing similar technologies helps establish not only existing works, but the assumptions made about how these works will be deployed.

The Permanente Group in California tested an AI system to perform the duties of a medical scribe by first transcribing patient-doctor conversations, then summarizing key points and generating patient notes [46, 47]. Tierney et al. found that the AI scribe, when generating unstructured notes, reduced clinicians' time spent in the records system outside of normal working hours, and time spent taking notes during patient visits. In post-study interviews, both patients and clinicians found that the reduction of typing during patient visits improved the interpersonal relationships [46, 47], *allowing the clinician to spend more time on medicine.* Smaller-scale evaluations of the Microsoft Dragon Ambient eXperience healthcare AI assistant [14] as a scribe found mixed results, with Liu et al. finding no statistically significant change in physician efficiency across a longitudinal study [29], while Cao et al. found an improvement in efficiency during their pilot [9], highlighting the importance of setting to the success of an real-world AI deployment.

Finally, industry applications also motivate the use of LLMs to reduce data entry burden, with tools such as Google Meet embedding LLM notetakers and summarizers into virtual meeting rooms [19].

There is a common desire to let the AI system to take over the banalities of paperwork so users can focus on applying their specialized skills. However, while similarities exist between the conservation setting and hospitals or meeting rooms, prior work assumes internet-equipped environments where applications can send API requests to speech-to-text or large language models hosted on cloud-based GPUs [14, 19, 32], and do not consider on-device deployment as a possibility. For many field staff working without cellular service, on-device inference is the only way to achieve real-time transcription and summarization. Further, in many of the previously studied application arenas, data entry consumes a high percentage of the user's work time [45], which may not be the case in conservation.

2.3 Technology Acceptance Theory

Davis [12] theorizes that the core elements which make a technology acceptable to a given user are its perceived *ease-of-use* and *usefulness.* As discussed previously, conservationists place high demands upon the hardware and software they allow into their toolbelts. The *usefulness* of a technology, here, means that it must do the job *and* last under extreme conditions. Given staffing and training shortages as well as varied literacy across users [31], it is critical that, as Davis posits, new technology is seen as *easy-to-use* in conservation.

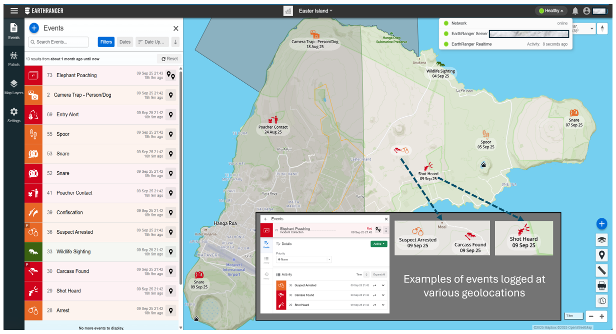
Subsequent work by Venkatesh et al. expands to include external factors that may influence the *behavioral intention* of a user towards a new technology and thus, the rate of user adoption [49]. They identify four constructs that influence a user's behavioral intention towards a technology: *performance expectancy, effort expectancy, social influence,* and *facilitating conditions* [49]. This Unified Theory of Acceptance and Use of Technology (UTAUT) is particularly useful for explaining acceptance of conservation technology because it is able to encompass hierarchical structures and infrastructures, beyond Davis's more internally-facing and individual perceptions.

Davis and Venkatesh's works provide a theoretical foundation which allow us to understand where incorporating AI — in particular customized on-device LLMs — into existing systems and workflows may succeed, and where it may fail.

2.4 Speculative Methods

To answer the question of how users will perceive the integration of on-device AI into their existing workflows, we need to consider their current infrastructure. However, it is both risky and impractical to fully deploy highly speculative technologies, and as such, we lean on design methods to envision how our proposed technologies might improve or disrupt user systems.

Thus, we turn to speculative design, which allows us to understand whether we are targeting the "right" problems when designing systems and challenge prevailing assumptions about certain technologies [54], which we find useful when considering applications from a rapidly-expanding and optimistic field, such as LLM research. For this work in particular, given its intention to understand how a newer technology with **significantly higher infrastructural requirements** compared to the existing technologies used in conservation, we drew from "infrastructural speculations" as proposed by Wong et al. to think beyond the direct users and into the broader network of support and infrastructure assumed to exist by both cloud and on-device LLM deployments [53]. The higher cost and volume of computing resources required to support AI, especially AI training, make infrastructure a core focus when considering the holistic impact of an AI deployment.



13 results from **about 1 month ago until now**

73	Elephant Poaching	09 Sep 25 21:43 18h 8m ago
2	Camera Trap - Person/Dog	09 Sep 25 21:42 18h 9m ago
69	Entry Alert	09 Sep 25 21:42 18h 9m ago
55	Spoor	09 Sep 25 21:42 18h 9m ago
53	Snare	09 Sep 25 21:42 18h 9m ago
52	Snare	09 Sep 25 21:42 18h 9m ago
41	Poacher Contact	09 Sep 25 21:42 18h 9m ago
39	Confiscation	09 Sep 25 21:42 18h 9m ago
36	Suspect Arrested	09 Sep 25 21:42 18h 9m ago
33	Wildlife Sighting	09 Sep 25 21:42 18h 9m ago
30	Carcass Found	09 Sep 25 21:42 18h 9m ago
29	Shot Heard	09 Sep 25 21:42 18h 9m ago
28	Arrest	09 Sep 25 21:42 18h 9m ago

No more events to display.

Events
73 Elephant Poaching
Incident Collection
Red
09 Sep 25 21:43
Details
Active
Priority
None
Activity
36 Suspect Arrested 09 Sep 25 21:42
30 Carcass Found 09 Sep 25 21:42
29 Shot Heard 09 Sep 25 21:42

Examples of events logged at various geolocations

Figure 2: Screen capture of the EarthRanger desktop application. On the right is an example map, displaying geolocations of various logged events. On the left is a list of recent events.

Infrastructural speculation allows us to center the *broader system around a technology*, as opposed to the technology itself. As Wong et al. suggest, the broader system can include regulatory frameworks, energy and hardware requirements, and maintenance accessibility of a technology [53]. As a field, conservation often occurs transnationally in settings where the expectations of the Western technology majors may not apply. Politics and policy, resource limitations, and even physical and biological landscapes may have significant effects on technology acceptance and efficacy relative to studies taking place in wealthy developed countries.

3 Background

Here, we describe our goals, our partners, and our technology for this project.

3.1 Community partners

In the following section, we introduce the EarthRanger platform and our community partners, Panthera and the Giraffe Conservation Foundation (GCF). Panthera and GCF both employ the ER application for staff to record data in the field. We introduce the field contexts wherein staff log observations, which we aim to support using on-device LLMs.

3.1.1 EarthRanger. A large and critical part of the work of conservationists, ecologists, and other scientists is gathering and understanding real-time data from many sensors, individuals, and animals across broad geographic areas. *EarthRanger* (ER) [51] provides a real-time web and mobile platform to aggregate information, from the locations of collared elephants to ranger patrol routes. EarthRanger consists of GIS-based tools for conservancies to meet a number of their critical needs, including (1) assisting in the tracking and monitoring of wildlife, (2) capture of human-wildlife conflict, both accidental (e.g., wildlife entering human areas) and intentional (e.g., poachers), and (3) coordinating rangers—conservancy field staff who enter the conservation area to protect and manage the animals. ER aims to bridge the gap between the vast volumes of data arriving from disparate data sources and the limited bandwidth of workers at partner organizations to prioritize and act upon the information. Figure 2 shows an example of the ER interface.

ER operates in a vast array of contexts around the world, from the classic "wildlife conservancy" with fences and protected land, to oceangoing shark trackers, to field staff who monitor national parks. Operating on a client-service system, it provides not only a core platform but also technical support and **customization** to each of their hundred-plus partner conservation organizations; this is necessary, as many do not have a software engineer on-staff. ER's service infrastructure, shown in Figure 3, allows us to envision extensions to the core application that support partner-specific customizations instead of one globally generalizable application.

The specific functionality we focus on in this project is their "Event" form, which allows users in the field to summarize their field observations as highly-structured data. We found that organizations often use the form to track animal populations and movements, or to record specific qualitative observations of animal habitats. Event forms function both on- and offline, as does much of the ER app.

3.1.2 Panthera. Panthera is a charitable organization founded in 2006, dedicated solely to conserving the world's 40 wild cat species and their ecosystems by developing strategies for protecting both large and small wild cat species. Their work spans 35 countries across North and South America, Asia, the Middle East, Europe, and Africa. We worked specifically with Panthera's Olympic Cougar Project (referred to as Panthera), which seeks to protect and connect cougar populations on Washington's mountainous Olympic Peninsula, and has worked on mapping wildlife corridors, identifying barriers to movement, and collaborating with six Tribal nations to improve genetic diversity and long-term viability.

Panthera has fitted dozens of cougars in the Olympic Peninsula with GPS-enabled tracking collars; the data from these collars are aggregated on the ER platform — the collars ping the ER servers hourly, transmitting the cat's location. The platform will then identify notable "clusters" based on certain cougar behaviors, such as bedding down or eating a kill [52].

3.1.3 Giraffe Conservation Foundation. GCF is headquartered in Windhoek, Namibia, and operates across the continent of Africa, conducting scientific research, organizing community outreach and education programs, and raising funds for giraffe conservation. It was founded in 2009, but its roots hearken back to the late 1990s, a time when little research had been produced about the giraffe. Using science-based conservation to save and study giraffe in the wild is the core of GCF's mission, and they work across Africa in a number of areas, from illegal trade prevention to translocation of giraffe populations.

Like Panthera, as part of their resesearch agenda, GCF will tag giraffes and track their movements with ER; they also conduct driving surveys of both tagged and untagged giraffe populations. ER serves as both a data management and location visualization system for them.

3.2 Project vision

In early conversations with EarthRanger (ER) around novel technologies they were interested in exploring, we discussed using LLMs as on-device assistants to help improve the data collection for field staff using ER. From there, we chose to focus on exploring how speech-based language model assistants would fit into the complex ecosystem of conservation field work, given the significant constraints imposed by the great outdoors on conservation technology. We endeavored not to imagine a system that would replace ER users, but rather facilitate spending more time employing their naturalist skills instead of doing data entry.

3.3 On-device LLMs

We briefly introduce the current state of on-device LLMs and contextualize them within this study. On-device LLMs could not only enable internet-free processing, but also safeguard user privacy and improve user control relative to cloud LLMs. For example, fast-moving cloud LLM API providers may suddenly discontinue older models, which may not align with the slower pace of conservation technology development. And when working across countries with varied data protection and governance regulations, on-device models reduce the policy barriers to technology adoption by keeping sensitive user audio recordings and protected animal data out of the cloud. If leaked to poachers or other bad actors, audio can be used to identify locals collaborating with conservationists; one study found that geolocation of audio samples using background noise is possible [10]. Edge processing also reduces the amount of data that potentially has to be stored — instead of storing and transmitting audio data, only the extracted data entities needs to be synced with centralized databases. This also reduces the cost of compute, as on-cloud hosting generally incurs either a per-call cost, or the cost of self-hosting models. In contrast, open-source on-device models can be used for free.

Most language models "small" enough to run on mobile devices—Llama 3.2-1B [18], Qwen3-0.6 and 1.7B [56], Phi-4-mini [30], to name a few—lag behind their larger counterparts in performance *on general tasks*: Llama 3.2-**1B**, for example, averages 14.4% accuracy across OpenLLM benchmarks, while the earlier Llama 3.1-**70B** achieves 43.41% accuracy [15]. Here, we consider "small" to be a model runnable on commodity mobile devices, < 4B parameters, as found by Yan et al. in their evaluation of various LLMs on mobile devices [55]. Models running on-device also exhibit significant latency relative to their cloud-hosted peers [55], as most mobile devices lack the capacity for massively parallelized matrix multiplication.

Once they have undergone supervised fine-tuning (SFT), small LLMs have shown themselves capable of matching (if not exceeding) the accuracy of larger models *on specific tasks*, such as sentiment analysis [7, 23, 57]. Some argue that the tunability, relative efficiency, and general operational suitability of small LLMs over true LLMs makes them a better fit for agentic and assistant systems [3, 23].

Regardless, for our purposes, supposing little to no internet access in the field, any LLM used would have to run on a mobile device, and thus would have to be small. While many important studies explore the systems and architecture optimizations necessary for on-device deployment of LLMs, very few implement user-centered LLM applications locally. For example, in their journal design, Moëll

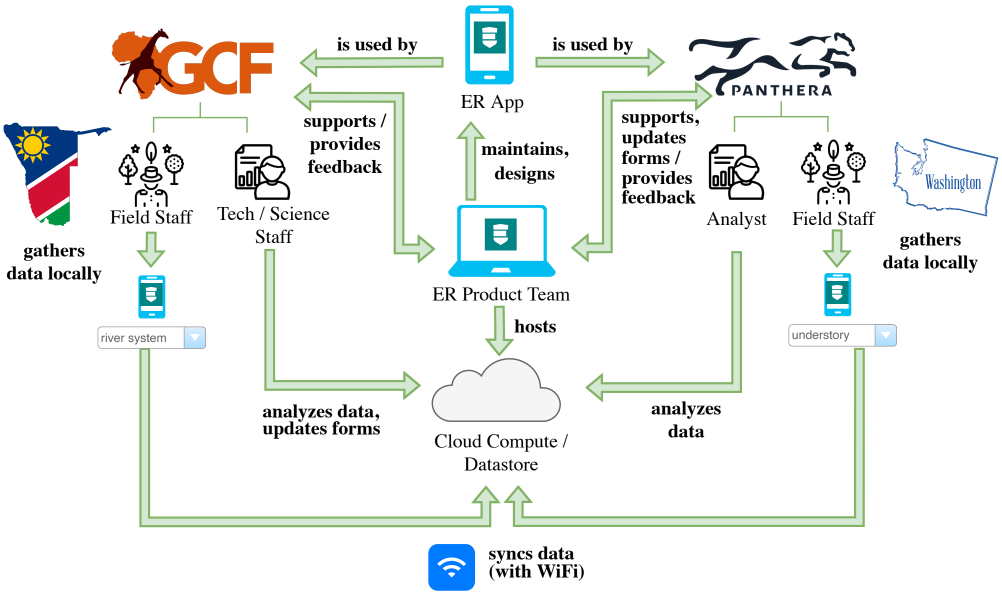


Figure 3: Diagram of the interaction between EarthRanger (ER) and two of its many partner conservancies, showing both how ER provides services and how its partners use ER and provide feedback.

et al. include an option for information to be processed by a local LLM, but do not deploy this functionality, instead relying on cloud-hosted APIs [32].

3.4 Ethics

Our studies were conducted under approval from our institutional IRB, who found them to be exempt due to the low risk to participants. However, we still gathered consent from participants. Before recording data, we made clear to all participants that we were conducting user research around on-device artificial intelligence in collaboration with EarthRanger, and aimed to publish our findings. We engaged with the users by observing and participating in their daily work through their employers. With our participant observation approach, we did not increase their risk exposure beyond levels inherent to their jobs. We also did not provide compensation, as we were mostly observing their daily activities or asking them to slightly modify their data entry process. All personally identifiable data, including audio recordings, were stored in secured accounts. All training was conducted on an on-premise GPU; no data was sent to cloud LLMs.

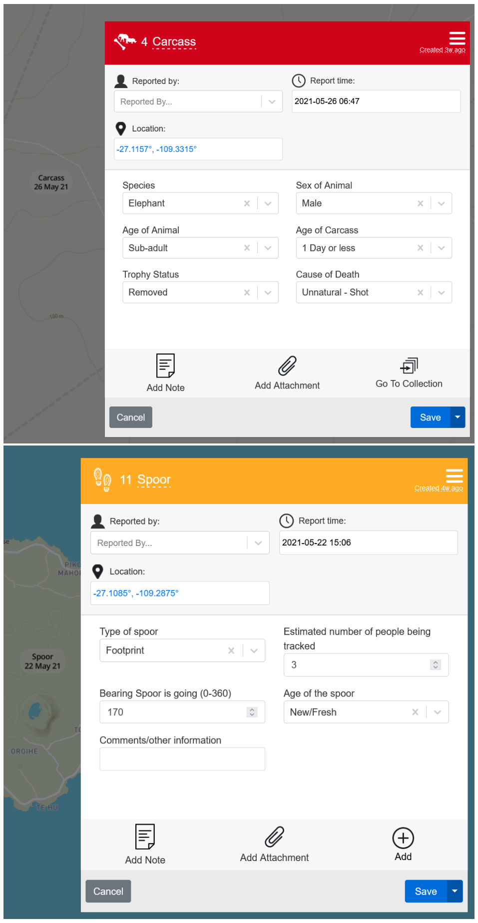
4 User Investigation: Methods, Field Context, and Findings

To better understand how field staff perform data entry, we conducted participant observation with Panthera and GCF in their work environments. For more user perspectives, we also issued a survey, with the help of ER, on current usage of the ER Event form and thoughts on audio as a user interface. In this section, we lay out the qualitative methodologies we used to understand their respective working contexts, and the field work we observed while visiting our partners. We also describe how the infrastructure ER provides supports the suitability and maintainability of the application.

4.1 Methods: Participant Observation and Surveys

Participant Observation. Participant observation places the research team amidst the study subjects, allowing the researchers to observe and interact with workers directly in their working environment [35]. Participant observation traditionally involves four (sometimes co-ocurring) stages: (1) establishing rapport, (2) immersing oneself in the field, (3) recording data and observations, and (4) consolidating the information gathered. For each of the teams, (1) was an ongoing endeavour conducted through online discussions, a shared meal, a long drive to the field site, and continuing communication. We then completed (2) and (3) at the field site, and (4) once we had returned from the field work, with all stages repeating in various ways as we worked with different organization. Participating directly in-situ allowed us a deeper, multisensory understanding of how Event forms are used.

For this study, we joined field staff at Panthera and GCF on field work in which they used the ER application for data entry tasks and learned how to observe and enter data ourselves. We spent one day in the field with Panthera in the Washington Olympic Peninsula in May of 2025, and five to six days in the field with GCF in Windhoek and Etosha Heights Game Reserve in August of 2025. During these sessions, we recorded audio of various tasks and informal interviews conducted in the field. With Panthera, we observed recordings of Bed and Kill Site data across several locations. With GCF, we observed field staff on various animal



4 Carcass

Reported by: Reported By...
Report time: 2021-05-26 06:47
Location: -27.1157°, -109.3315°
Species: Elephant
Sex of Animal: Male
Age of Animal: Sub-adult
Age of Carcass: 1 Day or less
Trophy Status: Removed
Cause of Death: Unnatural - Shot

11 Spoor

Reported by: Reported By...
Report time: 2021-05-22 15:06
Location: -27.1085°, -109.2875°
Type of spoor: Footprint
Estimated number of people being tracked: 3
Bearing Spoor is going (0-360): 170
Age of the spoor: New/Fresh
Comments/other information:

Figure 4: Filled-out EarthRanger Event forms representing information obtained from an observed animal carcass and spoor (footprints) sighting.

sightings, such as Elephant Herds; groundskeeping tasks such as Fence Break or Camera Trap; Patrols where routes were traced via GPS, and Giraffe Herd sightings.

Participants. The participants in our study (other than ourselves) were members of the field staff for the partner organizations working on projects that leverage the ER platform. At both Panthera and GCF, an individual project or geographic area of focus (*e.g.*, Olympic Peninsula or Etosha Heights) may only have a small number of associated employees, often drawn from the local community. For example, Panthera operates globally, but their Pacific Northwest cougar-counting initiative is conducted by a few paid staff, stationed locally; thus, our participant observation was with just two staff members. GCF follows a similar model: we shadowed the lead program officer for their EHGR project as well as the lead for their northwest Namibia project working together in Etosha Heights.

Information Analysis. The research team performed a thematic analysis of the qualitative interview data [6], an inductive approach that involves reading through data to identify patterns that can be generalized into more abstract themes. Three separate researchers reviewed transcripts of interviews and recordings from participant-observer field work and identified codes representing main themes throughout the artifacts. The researchers then met to align on their codes and come to a consensus on dominant themes.

Survey Study. For broader insight into the user base of ER Event forms, we requested that the ER user experience team distribute a lightweight Qualtrics survey to ER users, represented as a request from an external research group. Surveys are frequently used in HCI to elicit user feedback, understand user characteristics, and usage patterns [34]. We translated the survey to ER's most commonly-used languages: English, Spanish, French, Portuguese, and Swahili. The survey contains a series of questions focused on usage of the form and user sentiment towards audio-based data entry. Questions are available in the appendix.

We received seven responses out of the 100+ organizations which actively use ER. One respondent was a new user who had begun testing ER for their organization, but had not yet used it in the field. The other six respondents spanned multiple continents and languages: one response was received from North America, four from Africa, one from Central America, and one from South America. Four primarily use English when dealing with ER, one uses Spanish, and one uses French. Four use ER Event forms daily, one uses them weekly, one uses them a few times a month.

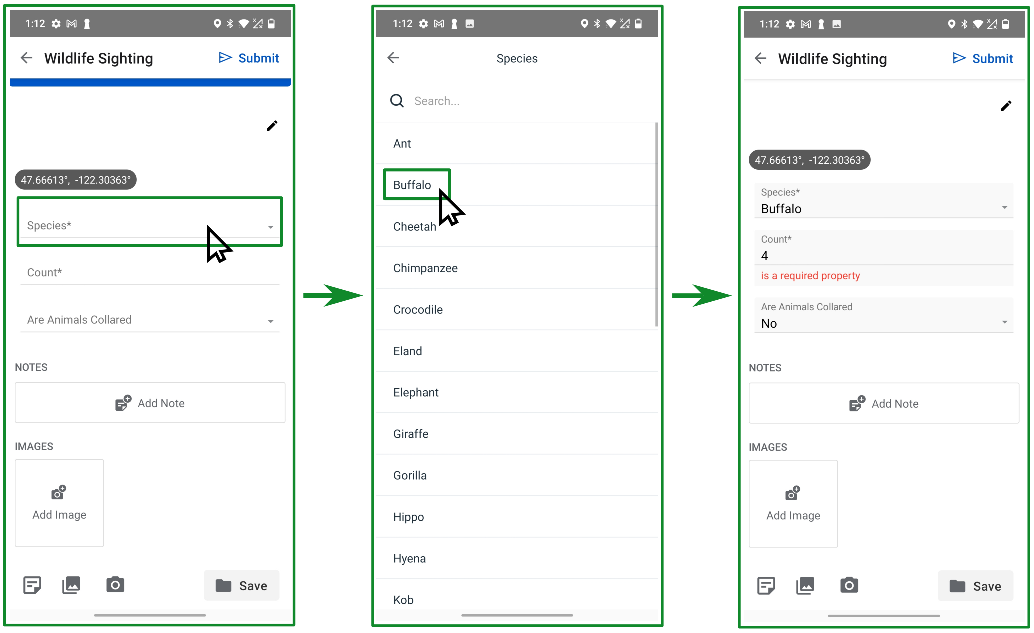
4.2 Field Context

In this section, we describe the field work undertaken as part of our participant observation sessions. These descriptions provide context for how ER is used in the field, and inform our speculative methods described in Section 5.

4.2.1 Panthera. We spent a day with Panthera at a number of sites on the Olympic Peninsula—a mountainous, forested region of the American Pacific Northwest which runs along the western coast of the continent. Panthera field staff most commonly travel alone, but had arranged for two people to accompany us to a number of bed and kill sites, where cougars sleep and consume prey, respectively. Panthera visits these sites to collect data on cougar activity in specific environments while also assessing the general environmental health of the ecosystems where the cougars live.

The day began for the field staff with the loading of "clusters" of cougar geolocation points onto GPS systems; these clusters are identified by ER as kill or bed sites based on the animal's movements, and interesting clusters are hand-selected for investigation. After we met up with the field staff, we drove up into the fir forests, to the closest road-accessible location to a kill site, at which point we dismounted and continued on foot. The field staff navigated with a handheld GPS system, which they preload with the ER-generated coordinates for the cougar point cluster before each outing. Panthera also employs onX, an offline-capable mapping application.

At each site, we hiked off-road, following the GPS system. Some of the sites we visited required hiking through dense brush, up and down steep, uncleared inclines, which is typical terrain in the region. Once we reached the destination, the field staff closely inspected the woods for signs of cougar activity — prints in the mud, hair, scavenger tracks, and prey carcasses. At some points,



Species* Buffalo

Count* 4

is a required property

Are Animals Collared No

Figure 5: Example workflow for filling out an ER Event form on a mobile device. The user selects an event type, then selects a dropdown field. Once they select an option, the user fills out the form and submits the event.

they got down on their hands and knees to look for minute, wispy hairs to determine whether a gentle depression in the grass was the result of a resting deer or cougar. On that day, the forest was exceptionally quiet and calm, though one of the field staff told us that she has worked through storms and bad weather. We visited 4-5 sites.

The Panthera field staff were always very enthusiastic when showing us various species of plants and explaining various tidbits of naturalist knowledge to us while hiking, such as how to tell apart various cone-bearing trees, or determining the difference between deer hair and cougar fur. They also explained in detail how they fill out the “prey age” field in their forms, which involves assessing the shape of the teeth.

Once they confirmed a bed or kill site, the field staff logged their observations, as well as hyperlocal habitat information, using the ER Events form. We show an example workflow on a mobile device in Figure 5. The field staff may also attach images of the site, prey remains, or scat, taken with their phone cameras. Recording habitat information required them to identify the dominant plant species and estimate the age of the surrounding forest, for example. The process of form entry alone could take five minutes, even if they had already inspected their surroundings for the relevant information. The data from the forms may be analyzed at a later date for insights into cougar behavior and population dynamics.

4.2.2 GCF. One member of our team visited GCF at their Windhoek headquarters, as well as their program station on a private game reserve, Etosha Heights (EHGR). EHGR is about five hours north of Windhoek by car, and lies along the southern border of the much larger Etosha National Park. The landscape is a mix of plains and hills, with some trees interspersed amidst the savanna grass.

At EHGR and other program sites, GCF uses ER as part of their giraffe surveys: driving along a predetermined, replicable route, field conservationists will visually scan for giraffe herds. ER is used to track their progress and note any deviations from the route, in addition to noting down specific giraffe which they spot. In general, they will travel with at least two people in the car.

One member of our team went out on two different types of field outings with GCF over the course of five days: two giraffe surveys and multiple trips to replace camera trap equipment. The giraffe surveys involved anywhere from 6-9 hours in the car, driving 100+ kilometers. The researcher went out with a different GCF field staffer on each survey, as well as one volunteer or supporter visiting GCF.

The giraffe surveys begin early in the morning to avoid afternoon heat. Once the team reaches the survey route, the Patrol function on ER is engaged, logging the GPS location of the ER device every few minutes. All members of the team begin looking out of the window of the Land Cruiser for giraffe. As shown in Figure 1, once someone spots a giraffe (Fig. 1a), one person will begin taking photos (Fig. 1b), trying to get an image for each side of the giraffe. The other person in the car will operate the ER app, filling out a form for each individual giraffe with its age, sex, and the camera number of the photograph corresponding to the giraffe’s left and right sides, respectively (Fig. 1c). Sometimes, it may be impossible to capture both sides (or even a single side) due to positioning, movement, or thick brush; in those situations, the form is only partially completed. Giraffe skin patterns are comparable to fingerprints, and GCF keeps digital ID books with photographs of the left and right sides of individual giraffe. After a survey, and *only once they have sufficient service or WiFi*, the field conservationists will sync the ER Events to the cloud, possibly manually inspect the new data for any giraffe they can ID by sight, then send the data to GiraffeSpotter [4], a software which automates individual identification based on skin pattern. As with Panthera, data is analyzed for various questions and projects.

At multiple points during each of the surveys, the GCF field staffer would climb on top of the vehicle, make sounds at the giraffe, or drive forwards/in reverse to get a better angle for photography. One of the field staffers used a HotWav Rugged tablet, and the other used their personal iPhone.

Aside from counting and photographing giraffe, the GCF field staff were always on the lookout for and enjoyed seeing many other species of animals, such as mountain zebra, various hawks and bustards, and elephants.

GCF also uses ER Event forms to monitor the containment fences which run between EHGR and the national park, livestock farms, and other game reserves. This involves driving dozens of kilometers along the fence and marking any broken areas (often trampled by elephants), which can take many hours.

4.3 Findings

From interviews, participant observation, and a survey, we were able to better understand field work conditions and the desires of field staff. From our learnings, we distill a set of systems and practices which make ER a suitable and maintainable conservation technology.

4.3.1 What makes the existing system deployable? In this section, we observe how the "maintainability" and "suitability" of EarthRanger and other technologies make them field-deployable.

Maintainability — User Support System. Usability and maintainability are crucial, especially given that much conservation technology is developed in wealthy western countries but used by individuals in less well-resourced, often rural areas [5, 27, 41]. In our interviews with conservation workers, we found that usability and maintainability are tightly linked to **user control and configurability**.

ER is not just a software provider: they also offer services — cloud hosting, but more importantly for user control, significant customization of what data is collected in Event forms. This may be through a superuser model, where a technically-trained individual within the conservancy, or at ER, can quickly create and edit custom Event forms for the field conservationists. From our survey, we found that two users update their forms yearly, one updates their forms about three times a year, and two update their forms monthly. One never updates their forms. These results indicate that some ER users do elect to change their forms with some frequency.

At GCF, different project owners have custom control over Event form fields and field types, enabling them to adjust the technology to fit their data collection needs, instead of the other way around. Users we spoke with were highly complimentary of the agency ER provides them: a user at GCF said, regarding the usability of the Event form, *"We set [the form] up how best it works for me in the field. So I have an idea, and then I give it to [the GCF EarthRanger coordinator], and then she sets it up the way I want it to work for me. How easy is collecting data? Is it easier to first get the sex, or the age? Or is it easier to first put in the photo number? Those specific details. And so it's set up in a way I want it to be. If I find it difficult it after a period of time, then I can always change it."*

Not only does user control over form fields improve usability, it also allows for the organization as a whole to take ownership over certain aspects of maintenance as their missions change over time. For example, Panthera was able to add new Event forms easily and quickly when starting a completely new project, with just a few email exchanges. In short, the *services and infrastructure* which ER provides to their end users are what fundamentally make ER maintainable and continuously usable for its users.

Suitability — functionality despite limited internet access. By nature, conservationists often work in wild, remote areas. We found, from our survey and participant observation sessions, that internet access in the field is sparse to nonexistent: five of seven survey respondents said that they use EarthRanger without an internet connection. While in the field with Panthera and GCF, we noted that there was patchy to no cell service at the observation sites in western Washington state and the Etosha Heights Game Reserve (EHGR). Field conservationists may not re-enter an area with reliable internet for days. From members of the GCF team, we learned that their work in rural Angola must be done without any internet access for the duration of the trip. While advances in satellite technologies, notably Starlink, may broaden internet access, in countries such as Namibia and South Africa, Starlink has been banned for regulatory reasons [36, 37]. ER anticipates the intermittent service problem by enabling users to collect and store data offline. Once the user has service, they sync the data to an ER-hosted cloud instance, where others in their organization can view and export the data for analysis [51].

We asked users to imagine recording audio alone in-the-field, then waiting for *processing* by a cloud-hosted AI model once they had returned to service. Several said they like to check their data in-situ to ensure accuracy and completeness *while still immersed in the directly observable context of the field.* For example, one GCF user said that he would find it difficult to remember the small details of the giraffe which help him catch mistakes in the data entry; waiting would be extra work, cumbersome, and error-prone, especially given his long deployments without internet. Panthera users pointed out that it might be difficult to catch missed details if they don't see the form populate while still in the field, and hiking back out to the site to re-gather the information would be impractical.

Suitability — Simplicity. We observed that underlying the importance users place on reducing paperwork time, is a strong preference for working in the field over working at the office. Users at both Panthera and GCF have significant field expertise, but are often quite busy with other tasks. One user, when asked what the best part of their job is, said that it is *"the excitement of what you may see that day"*, in the field. As such, conservation technologies which increase the proportion of time conservationists spend doing what they enjoy most about their jobs seems beneficial to users.

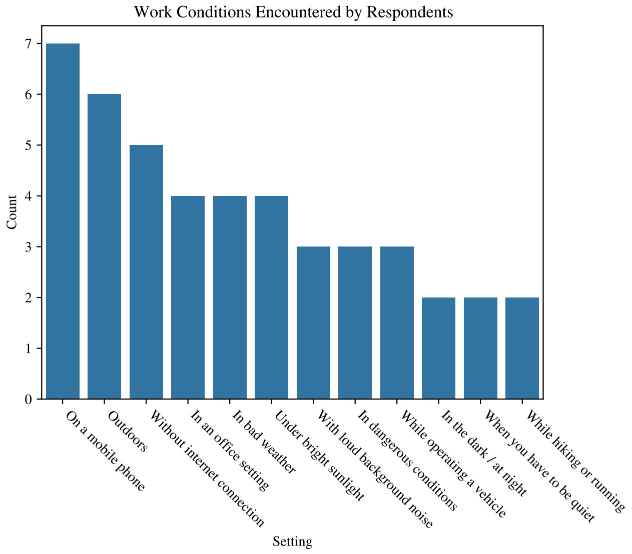


Figure 6: Bar chart representing survey responses on the type of conditions under which the EarthRanger application is used in the field. There were seven total respondents.

In addition, conservationists juggle many things at once in the field: operating a plane or an all-terrain vehicle, or watching for falling tree limbs and incoming elephants. Their jobs are highly complex and require specialized knowledge, and as such, having a quick, easy, and low-error form of data entry is paramount. Good conservation technologies should minimize time spent focusing on the technology itself, and make it easy to collect clean data. In the words of one GCF user, *"Don't collect shitty data. It will bite you in the bum."*

Some ER users told us that the simplicity of the interface is very important to them — survey respondents called it "quick", "no-fuss", and, in short, "simple". In our usage of ER during participant observation, we found the interface to be straightforward and intuitive. With the notable exception of Panthera, other organizations (Osa Conservancy, Save the Elephants, GCF, and the Smithsonian Zoo) had short forms (less than 7 fields) which required little typing.

4.3.2 Survey Responses. Survey responses highlight the diversity of conditions in which users work (shown in Figure 4). Outdoors and mobile usage are high, as is usage without internet. Bad weather and bright sunlight, which make touchscreen-based interfaces difficult to use, are common to more than half of users. Fewer users report using Event forms under more trying circumstances, such as while moving under their own power, or while operating a vehicle.

One focus of our survey which is not encapsulated by our other findings was how users felt when envisioning using an audio-based system for the ER Event forms. On this subject, we received mixed responses: five of seven respondents answered the question, and two of the five were somewhat uninterested. One said that they *"don't feel that the small amount of data to be entered on Events is worth the effort for voice to text functionality"*; another worries that audio would be more difficult to work with than their current multiple-choice system, which would be problematic as *"speed and convenience are key"*. This respondent pointed out that audio interfaces can be error-prone, especially in windy conditions, which could frustrate their volunteer users. Another respondent stated that they might be interested in having audio as another option alongside text entry. Two of the respondents were more enthusiastic, with one stating that *"sometimes the conditions do not favor entering data by hand"*, and another simply saying "great". In short, the respondents were open to the idea of using audio, but only if it was better than their existing system.

The users' satisfaction with Event forms as they currently exist may be part of why they are somewhat disinterested in introducing a novel system. On a scale of (extremely dissatisfied, somewhat dissatisfied, neutral, somewhat satisfied, extremely satisfied), six of seven users reported that they were *at least* somewhat satisfied with ER Event forms, with the seventh (being the user who was new to ER) saying that they were neither satisfied nor dissatisfied. One user stated that they would like to see form field formats be more "customizable, (e.g. box to tick, likert scale type button)." Other users suggested that improvements could be made to the syncing or GPS tracking accuracy, but overall, users found that the current Event forms work well for them in the field.

5 Technical Validation

Now that we have provided a foundation for understanding the ecosystem, we will discuss our technical probe. We explore the technical challenges of creating on-device AI, and use speculative design to imagine how said technical challenges will affect both the field staff and the systems which exist to support their technologies.

5.1 System Development

We set out to build a system that would help enter data on Event forms (Fig. 4) from speech. We settled on the implementation of a transcription model to transcribe the user-generated audio, and an LLM to extract relevant information and reformat the audio transcript as a JSON object, as shown in Figure 7. The JSON object would then be propagated, filling out the form fields automatically. We implemented model inference pipelines in the React Native framework to faithfully understand performance.

Data-gathering and Informal Interviews. We conducted introductory interviews with EarthRanger's software team, Panthera, GCF, Save the Elephants (Kenya), Osa Conservation (Costa Rica), and the Smithsonian Zoo to gain a foundational understanding of how the ER application is used. These organizations also provided us with various data artifacts which we could use as examples of Event forms.

Speech-to-Text Model. We began with the speech-to-text model, for transcription. From our tests of different devices, some smartphones, such as iPhones running iOS 18 and later, or Google Pixels, have built-in transcription services; for those that do not, we would have to provide an on-device transcription service. We elected to use OpenAI's open-source transcription model, Whisper [42], as its autoregressive nature allows for prompt-based "guidance" of the grammar. Prompting allows for training-free flexibility when dealing with uncommon data distributions, which we anticipated we

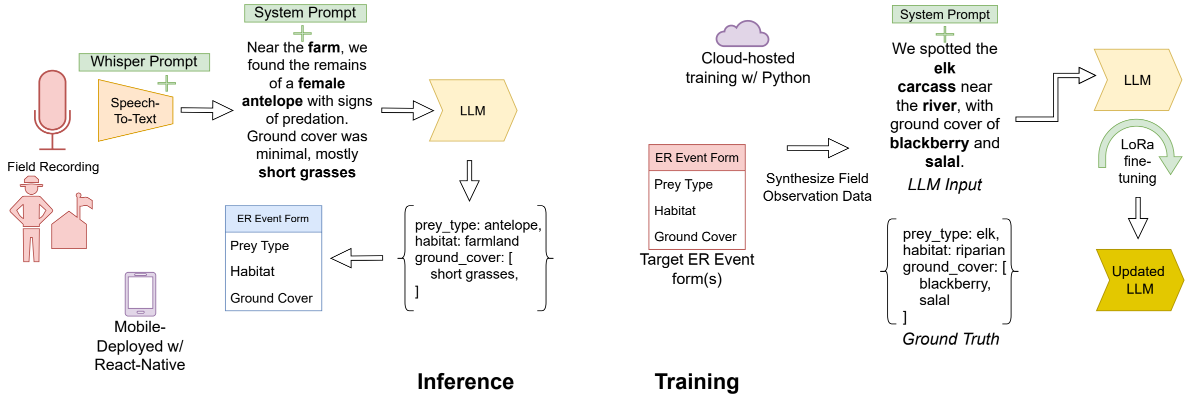


Figure 7: Diagram of the inference and training processes for our prototype transcription plus LLM pipeline. Left: Inference pipeline, where the user records an observation in the field; the audio is recorded and transcribed by a speech-to-text model. The transcription is then sent to the language model, generating the formatted JSON. Right: training pipeline, where a target ER form is used to synthesize dataset consisting of (audio, JSON) pairs; an LLM is then fine-tuned using parameter-efficient methods, resulting in an updated, custom LLM.

would encounter in jargon-heavy field observations. For on-device inference without GPUs, we chose to use whisper.cpp [17], a highly efficient C++ implementation library of Whisper models. We tested the small (466 MiB), base (142 MiB), and tiny (75 MiB) sizes, as well as the five- and eight-bit quantized counterparts of these models.

For each separate conservancy, we primed the model with a custom prompt, which consisted of a sentence with qualitatively selected uncommon words, jargon, and anticipated place names, such as "salal", or"Hoanib River".

JSON formatting. After the user's audio is transcribed, we submit the generated transcript to an on-device LLM, which, as shown in Figure 7, extracts from the transcript a list of values corresponding to form fields in the format "field: value", where "value" may correspond to a string, number, array, or object.

We impose an interface upon the LLM's outputs to ensure that they align with the data type, format, and selectable choices available in the corresponding form. When fields or values do not match exactly to the form fields or selectable options provided in the form definition, we use the Fuse.js library [44] to find fuzzy matches from a known set of field and value options.

LLM Fine-Tuning. We used Llama 3.2 1B [18], a transformer-based [48], open-source LLM. To fine-tune the Llama 3.2 1B model to take in speech-like text and extract form fields and corresponding values, we trained on a synthetically generated dataset. We fine-tuned the model to cover two of Panthera's event forms (kill site, bed site) and two of GCF's (giraffe mortality, giraffe sighting).

The datasets were synthesized using provided JSON templates for each form type. The system prompt was generated using the form fields and their descriptions. For each data instance, we either selected or generated a reasonable random value for each field. The ground truth label was formatted as a list of "field: value" pairs. The user query was generated by wrapping the field name and value in a sentence similar to an audio dictation. The sentences and "field: value" pairs were jointly shuffled to ensure order invariance.

Low-Rank Adaptation (LoRA) [20] was applied to the Llama 3.2 1B model, enabling parameter-efficient fine-tuning with the PEFT library. We fine-tuned the model until the validation loss plateaued. We then compiled the model in GPT-Generated Unified Format (GGUF).

5.2 Technical Assessment

Having generated a piece of software that could be used and tested, we gained a firmer understanding of the capabilities and shortcomings of our on-device LLM, made real.

Whisper Accuracy. We did not fine-tune the Whisper model, instead opting to prompt it with a sentence containing commonly-stated words and formats. As shown in Table 1, we tested Whisper-Base across about an hour of recordings from three different english accents (Namibian-German, Namibian, and American), four people, and four use cases. Recordings were taken by field staff themselves, with American English recorded in-the-field (with background noise). Error was calculated with jiwer's Word Error Rate (WER) package. The error here is estimated generously, as numbers (eight versus 8, for example) often arose as a source of error, and our ground-truth transcripts are imperfect. We found that though WER for certain accents and tasks may appear high, the downstream LLM can be fine-tuned to adapt to common mis-transcriptions.

LLM Accuracy. We collected rough metrics on the accuracy of our system for Panthera's Bed Site and Kill Site form, as it was the most challenging given its high number of field and field values, shown in Table 2. The test was conducted with synthetically generated transcripts held out during the training process as a separate evaluation set. Accuracy was measured after LLM extraction and subsequent fuzzy searching for close-enough values.

Accent	Subject	Total Recording Time	WER
Namibian-German English	Giraffe Mortality	10 min	10.5%
Namibian English	Giraffe Sightings/Mortality	27 min	24.1%
American English	Cougar Bed & Kill Sites	36 min	19.6%

Table 1: Word error rate (WER) by accent, using the base Whisper model. WER was calculated over transcripts joined into a singular paragraph.

Incorrect Fields	Missing Fields	Total Number of Fields
7 (1.3%)	19 (3.6%)	533

Table 2: From results extracted using a fine-tuned LLM, the number and percentage of fields missed due to either incorrect or missing values across 20 Panthera records.

We can see that overall, the LLM is able to extract fields with minimal error, even on the most challenging form we encountered across users. The rate of missing fields is higher than that of generating the incorrect value for fields, which can be advantageous — it is visually easier to notice an empty box than to read through existing text and look for possibly incorrect entries.

Upon inspection, many of the incorrect fields were a result of the model generating "no" instead of "unknown" or vice versa (3 instances), slightly incorrect latitude/longitude values (2), and slight misspelling or misformatting of the output (2). These errors could possibly be mitigated through further fine-tuning or adjustment of thresholds on the fuzzy search.

Systems Performance. To evaluate the system load of our on-device models, we obtained three devices (specifications shown in Table 3) based on some of the most commonly-used phone brands across EarthRanger users. Using these phones and Android Debug Bridge's **bugreport** functionality, we recorded fine-grained changes in battery level and internal CPU temperature for two modes of data entry: the existing method of typing in data by hand, and our new proposed method of audio plus LLM. For both modes, we completed both Panthera's Bed Site and Kill Site forms and opened the camera to attach a photo.

As shown in Figure 8, the existing method of data entry generally consumes less battery than using our method, while also reaching lower peak temperatures. These results are to be expected: as the size and nature of matrix multiplications required for AI models tax CPUs, especially on mobile, the systems capacity of the varied devices used to access the ER app become the limiting factor for model capacity and speed. We found that in using the transcription and LLM functionalities of our prototype, mobile devices such as the Google Pixel and the OnePlus drained their battery faster and became noticeably warmer than under normal operating circumstances.

6 Discussion

Returning to **RQ1**, through gathering context across participant observation, surveys, and user interviews, we established **user control**, **offline functionality**, and **simplicity** as core to the suitability and maintainability which make conservation technologies long-term field-deployable. Through participant observation and systems evaluation, we also addressed **RQ2**, finding that while on-device audio models and LLMs may provide opportunities to make data collection physically easier, the training and computational requirements of on-device LLMs in particular could be disruptive to existing EarthRanger deployments.

We now discuss the significance of these findings, applying speculative methods and technology acceptance ideas to imagine both the promise and the pitfalls of on-device LLMs, while also outlining some of the limitations and ethical considerations of our study. We also discuss possible directions for future work.

6.1 Insights from Technology Acceptance

In Section 4.3.1, we described our findings that both Panthera and GCF have come to rely upon EarthRanger because of its *maintainability and suitability* for their use case. Examining the introduction of on-device AI into this system through the lens of Venkatesh's UTAUT model helps us further understand how our preliminary findings when testing a prototype might influence user acceptance of on-device AI, especially given the preexisting benefits users see from using EarthRanger (ER).

6.1.1 Performance Expectancy. First and foremost, the threshold for **performance expectancy** will be high; while users raised complaints with the existing ER system, they emphasized that they are unlikely to adopt a new technology unless it is equal to or better than their current system.

We have shown that our system achieves low error rates after fine-tuning, and would likely not lower accuracy. Users also expressed that if on-device AI could provide immediacy and guaranteed uptime in no-service settings, it would improve their ability to process data and correct mistakes while remaining situated in the context where the data was observed or gathered over cloud-based AI. However, the *systems* performance of running AI models on mobile phones could be concerning for usability.

Firstly, while many users at Panthera and GCF have access to a vehicle for charging, when we asked them to imagine having to charge their device constantly, several problems arose. One user from Panthera noted that she sometimes hikes hours to reach a site; having to worry about battery charge would be annoying at best, and dangerous at worst. GCF field staff said that currently, their devices will stay charged for the entire day when using ER in the field; having to unplug and plug in their phones for charging would be cumbersome: *"The cables and then the connections and the bumpy roads. I don't think it's a good idea,"* one GCF user said. We also observed that the GCF staff will climb around the roof of the car, hopping in and out, up and down in order to better position themselves around the giraffe.

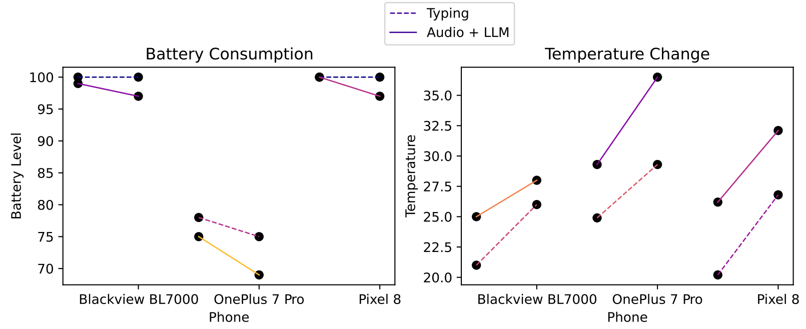


Figure 8: Battery consumption and temperature change comparison between Typing (existing method of data entry, dotted line) and Audio + LLM (our proposed method, solid line). From left-to-right, the first point in a line segment represents the average measurement around the start of the session, and the second point represents the average final measurement. Darker colors indicate greater net change in resource consumption over the course of one trial.

Device Name	Year	Processor	RAM	Time-per-Token (ms)
OnePlus 7 Pro	2019	Snapdragon 855	8GB	11.99±2.50
Pixel 8	2023	Google Tensor G3	8GB	8.65±2.22
Blackview BL7000	2025	MTK Dimensity 6300	8GB	6.93±0.49

Table 3: Specifications of test devices used.

Secondly, we noted when out with GCF, that device temperature can also be a problem. We visited Namibia towards the beginning of spring, when noon temperatures reached around 30°C; summertime temperature can rise to the mid-to-high 30s [40]. Already, we noted that the GCF user would move her tablet onto the dash so that it could get a more accurate GPS signal, then intermittently move it out of the sun when she realized it was hot to the touch.

Using our experiences from technical validation, we speculate that on-device AI deployments could reach concerningly high temperatures, especially in the sun or hot cars for day-long surveys. In turn, this could negatively impact the user's ability to record data if the device requires long cooldown periods, or if the device must remain plugged in.

6.1.2 *Effort Expectancy.*

How much more difficult would it be to use on-device AI in lieu of existing data entry? For our English-fluent and tech-savvy field staffers, all of whom had a Bachelor's degree, the new interface was easy to navigate, even if they did not understand how the underlying model was trained or deployed. Our system provided some ergonomic benefits — Panthera users said it would be useful for rainy conditions, for example, when precise typing becomes difficult on a smartphone.

We did not engage directly with organizations doing citizen science data entry. But from informal conversations with groups looking to have farmers or fishermen track wildlife sightings, we learned that low literacy, low familiarity with digital systems, and low language fluency pose significant barriers to digital data entry, especially when areas have little-to-no cellular service. We are hoping to expand our project in future work to investigate such cases, as we imagine that the audio-based modality of our system may lower the effort required to achieve user acceptance.

6.1.3 *Social Influence.*

Next, we turn to **social influence**, or the degree to which both conservation organizations and field staff specifically believe that others find it important that they use the system. There are three main social ties to focus on here: (1) the collaboration between EarthRanger and the conservation organization, (2) the internal structure of power between field and technical staff within an organization, and (3) external factors, such as trends amongst other conservationists or even society more broadly.

EarthRanger is highly responsive to feedback and is unlikely to push unwanted technology, especially given its nonprofit status. In organizations where field staff have a strong voice in what technologies they accept after testing, it is unlikely that leadership would force staff to use technologies that perform poorly, from our observations and interactions. Similarly, though it is difficult to forecast how society will adapt to AI systems, it seems unlikely, especially given the lack of direct competition amongst conservation organizations, that external pressures would significantly impact user acceptance of on-device LLMs.

Overall, we find that there is likely limited influence flowing through social ties at GCF and Panthera, and likely other conservation organizations. Conservation differs from the traditional corporate settings from which user acceptance models originate, and as such, social influence may be weaker.

6.1.4 *Facilitating Conditions.*

Finally, we assess **facilitating conditions**, or whether the user perceives that their organization has sufficient infrastructure, technical or otherwise, to support the deployment of on-device AI systems. We imagine our system in multiple scenarios, and from the perspective of not only direct users, but other auxiliary people, as suggested by Wong et al. [53], in order to de-center the technology itself and focus on the infrastructure. Specifically, we imagine the technology deployed in highly varied environmental settings not normally considered by developers, while also considering the broader network of support currently in place around EarthRanger. In doing so, we find one of the most significant hurdles for such a deployment: due to their small size and therefore small capacity, on-device LLMs must be customized to a task through fine-tuning. Existing infrastructure at conservation organizations does not seem well-equipped for the introduction of AI training for model customization into existing workflows, *especially* given that ER currently allows a great deal of user control.

To describe this problem in more detail, while fine-tuning significantly improved the on-device LLM's performance on a given task, when faced with inputs completely out of the fine-tuning distribution, the LLM cannot generalize. From the perspective of an ER superuser responsible for form updates, when receiving a request for a new field, instead of simply adding another line in a JSON object, the ER superuser would also have to synthesize new data, find sufficient GPU resources, run training, and re-deploy the model. Now, updating the form requires machine learning expertise and experience. From the perspective of an ER software engineer, training also requires more time spent on engineering and validation when compared to editing a JSON template, and the creation and monitoring of metrics post-deployment to ensure that the model does not drift in accuracy.

Expanding on the GPU hardware needed to reasonably train an LLM, we found that a mid-tier GPU (RTX 3090 Ti, 24 GiB of GPU Memory) was sufficient to do parameter-efficient fine-tuning (PEFT) with the 1 billion parameter LLama 3.2 model with careful memory management. In our speculative scenario, either EarthRanger or the conservation organization would have to manage and pay for such a resource, which can easily cost thousands of US dollars.

Some of our findings extend beyond the setting of conservation into broader user acceptance of on-device LLMs for low-latency agentic systems more generally — a growing area of discourse [3]. In particular, we highlight the accuracy of fine-tuned small LLMs; their enabling of users to stay immersed in a specific context, with or without service, regardless of whether any cloud model is available; and the possibility of increasing accessibility for low-literacy users. The potential problems we pointed out above also generalize beyond conservation — smaller, non-technical organizations (small businesses, local clinics, etc.) adopting AI will likely run into similar hardware and infrastructure problems when fine-tuning LLMs.

6.2 Ethics & Limitations

Here, we discuss ethical considerations of adopting our system. Given that our work is situated in ecological biodiversity and conservation, we must consider how our technology affects the environment. Luckily, our system does not require additional hardware or generate e-waste beyond what is already required to run EarthRanger — a smartphone. However, the power required for training, to run a single GPU per-model across multiple training iterations, and the resultant emissions, are potential downsides. Although lifecycle analysis is outside of the scope of our work, prior to actual deployment of our system, and indeed any AI system, the environmental cost of moving to more energy-intensive systems needs to be carefully weighed against benefits the system may provide. We note that conservation already struggles with the problem of emissions-versus-research tradeoffs, especially as they use many diesel-powered vehicles for offroad travel through rugged terrain or aquatic environments.

Another consideration we wish to mention is how LLMs may affect labor and employment. The core aim of our research was to understand and improve the experience of recording observations in-the-field and assess the potential for on-device AI. We did not seek to automate away or otherwise supplant field staff. In fact, from our study, we found that language models would be ill-equipped to handle the most critical and challenging aspect of these rangers' work: collaboration with local communities, interaction with animals, and mediating human-wildlife conflict. As such, we do not anticipate that our research will accelerate the automation of field staff jobs, though we acknowledge that the overall trend towards introduction of more technology into conservation may lead to unexpected consequences, especially as these technologies claim to improve productivity.

Finally, we would like to acknowledge that we interacted with a tiny subset of rangers and conservation organizations, limiting the generalizability of our findings. Following the understanding of generalization in Information and Communication Technology for Development (ICTD) inscribed by Burrell and Toyama [8], our goal is not to generate a general understanding of how on-device LLMs would work across all conservation organizations — such a claim would be difficult to make, given the extreme variability we observed across the organizations we spoke to. Even among our limited pool, we have met people who work from planes and people who work from zoos; people who track migratory megafauna and people who document illegally logged trees; the list goes on. Instead, our research aims to identify commonalities in order to address specific technical questions, such as infrastructural challenges to fine-tuning customized LLMs. We consider this work an initial exploration of a complex field, finding a place to start development of a technical "probe" [13, 21] rather than a broad exploration of an incredibly diverse and complicated technical ecosystem.

6.3 Emergent designs

One element that stood out to us from our field work was the enthusiasm and enjoyment all of the observed field staff gained from their work, especially in exercising their significant naturalist knowledge and encountering different plant and animal species. For example, an intern with GCF found great joy in keeping a species list of organisms she had encountered while at the reserve, and one of the GCF field staff said that the best thing about her job was the anticipation of what you would see that day.

Simultaneously, we also noted that GCF and Panthera collect different volumes of information: Panthera users record a plethora of habitat-related information, which may not make its way into any analysis. But as one Panthera user stated, “Let's say you hiked 3 miles into a cluster. You might as well get it all.” On the other hand, GCF collects very constrained data — predominantly location, giraffe age, sex, and photo numbers.

During a conversation with the Panthera users after showing them a demo of the prototype, one noted that it would be nice to use our system to record, transcribe, and somehow organize and publish his field notes for others to be able to look through: *"If I got hit by a car tomorrow...someone could go back and be like, what did [he] do this last decade? And it would mean something instead of it just being a bunch of notebooks that are totally incoherent, like I currently have in my life."*

From these observations, we imagine a speculative design for an ad hoc field journal for the rangers, allowing them to record not only the data required by ER Event forms, but also record things that bring them joy or keep their naturalist knowledge sharp. The data could even be shared and analysed using LLMs for patterns. For example, the GCF team was particularly interested in a strange, new giraffe disease which manifested as large lumps under the skin, but they lacked official mechanisms to track information around sightings. Field staff could use a local transcription model for quick note-taking; at the organizational level, a more powerful, cloud-based system which requires minimal personalization could be used to trace patterns in the data without requiring ER or the organization to support resource-intensive model training.

6.4 Cautionary Tales

The core goal of this paper is to understand how on-device AI might affect the work of data entry in the field. We have already hinted at a number of usability concerns: increased battery drain and overheating from AI applications; and a longer, more complicated, and error-prone iteration cycle standing between users and control over the form formats. In short, using the LLM would impose **high costs**.

Panthera have a very specific combination of factors that make our prototype design more useful for them — they spend lots of time scrolling through many options when entering data; they also work without assistance much of the time. For them, a cost-benefit analysis might end up in favor of the LLM. But for GCF, or some of the other organizations which we spoke to, who have much simpler forms, data entry is not the problem — the problem is getting a giraffe to face the right way, or getting the GPS to sync correctly.

Furthermore, we can envision a scenario where the on-device assistants are used not to enhance the joy of the work, but rather, to simply make field staff more efficient in the specific metric of data entry. Participant observation showed us that there is no AI that can replace having another person the sensory capabilities of another person tracking the animals alongside you as you call out information — and certainly none that can help you open the pickle jar for your field lunch. From our observations, there are certain intangible aspects of field work which field staff hold dear. For deployments of AI to improve the experience of the field staff, the AI must be well-aligned to the desires of the staff themselves.

We have proposed some promising speculative designs from our investigation, but we have also encountered the pitfalls suggested by others in the literature. We find that it is important, when faced with a rapidly growing technology that promises many benefits, to carefully consider whether the tech will be genuinely helpful — especially when working in contexts such as field work, which most western developers do not envision when designing products. We have seen in the literature a strong desire to avoid colonial patterns of power transfer between biodiverse and technology-producing countries; while on-device AI holds promise for conservation work, it must be deployed with care to avoid the potential pitfalls.

6.5 Future potential for on-device AI

Although we advise some caution, the future may yet be open to on-device AI in conservation work. Panthera's use case is a good candidate; the length and complexity of their forms and the single-person nature of their field work mean that our prototype could reduce the time they spend doing data entry. For organizations with similar circumstances — complex data, or data that must be entered while operating a moving vehicle or under bad weather conditions — we envision that on-device AI to transcribe and parse audio streams would be helpful.

While we did not directly evaluate community-based data collection, we had informal preliminary conversations with several conservation groups that would like to deploy form-based data entry with community groups living in close proximity to protected wildlife. However, such deployments currently face a major challenge — users may have limited digital or general literacy and speak a wide range of languages. Part of our exploration of on-device models is intended to establish a proof-of-concept platform that could enable future voice-to-voice form filling: **live speech-to-text, text translation and form-filling via LLM, followed by generated spoken reminders to recollect any missed fields**. All of the groups we spoke with said that such a workflow in remote areas would necessitate on-device AI due to both cost and network availability.

As we find methods of compression and pruning which may improve the efficiency of small LLMs, and as mobile devices become, on average, more powerful and in possession of greater memory, we may see that on-device AI becomes faster and less energy-hungry, making it even more advantageous in certain settings. However, as software and hardware develop to improve AI, we may see internet coverage broaden in parallel, especially with technologies like Starlink bringing access to rural areas, which complicates the cost-benefit analysis for on-device AI.

We have already begun the deployment of a version of our system with other organizations, and hope to collect and analyze this data to gain a deeper, user-centered view of how on-device AI will affect their workflow; continuing deployments with other organizations is a direct future work which we have planned.

Given our findings that LLMs constrained to mobile-friendly sizes necessitate fine-tuning and customization to perform sufficiently well, we believe that there is much research to do on the infrastructures which will have to support such levels of customization, as well.

7 Conclusion

In this work, we present our findings from time spent using ER with Panthera and GCF, two conservation organizations working in very different contexts. We sought to understand the potential of on-device AI to assist in data collection work in conservation.

From semi-structured interviews and participant observation conducted with field staff from both organizations, we were able to explore their work environments and identify the design decisions which contribute most heavily to the success of field-deployed technologies such as EarthRanger.

We then developed a system to establish the technical feasibility of on-device AI, in terms of both system performance and accuracy. We found through speculative methods that while current models can be deployed on-device with reasonable accuracy, the significantly greater compute required to both run and train such models could destabilize the systems and design choices currently in place to ensure conservation technologies remain suitable and maintainable.

Finally, we ideate on alternative designs of AI systems to improve the experience of field work for staff, and raise a few cautionary notes on the costs versus the benefits of deploying on-device AI systems, presenting avenues for future research.

Acknowledgments

This work was funded in part by a gift from Microsoft Research, and the Gaetano Borriello Endowed Fellowship for Change. The authors acknowledge Esther Jang, Innocent Obi for their help throughout the process, and Nicholas Day for compute resources. We also thank Jes Lefcourt and those at EarthRanger, Panthera, and GCF, who were instrumental in enabling the field work herein.

References

[1] Jorge A Ahumada, Eric Fegraus, Tanya Birch, Nicole Flores, Roland Kays, Timothy G O'Brien, Jonathan Palmer, Stephanie Schuttler, Jennifer Y Zhao, Walter Jetz, Margaret Kinnaird, Sayali Kulkarni, Arnaud Lyet, David Thau, Michelle Duong, Ruth Oliver, and Anthony Dancer. 2019. Wildlife Insights: A Platform to Maximize the Potential of Camera Trap and Other Passive Sensor Wildlife Data for the Planet. *Environmental Conservation* 47, 1 (Sept. 2019), 1–6. doi:10.1017/s0376892919000298

[2] Michael R. Appleton, Alexandre Courtiol, Lucy Emerton, James L. Slade, Andrew Tilker, Lauren C. Warr, Mónica Álvarez Malvido, James R. Barborak, Louise de Bruin, Rosalie Chapple, Jennifer C. Daltry, Nina P. Hadley, Christopher A. Jordan, François Rousset, Rohit Singh, Eleanor J. Sterling, Erin G. Wessling, and Barney Long. 2022. Protected area personnel and ranger numbers are insufficient to deliver global expectations. *Nature Sustainability* 5, 12 (Oct. 2022), 1100–1110. doi:10.1038/s41893-022-00970-0

[3] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. Small Language Models are the Future of Agentic AI. arXiv:2506.02153 [cs.AI] https://arxiv.org/abs/2506.02153

[4] Tanya Y. Berger-Wolf, Daniel I. Rubenstein, Charles V. Stewart, Jason A. Holmberg, Jason Parham, Sreejith Menon, Jonathan Crall, Jon Van Oast, Emre Kiciman, and Lucas Joppa. 2017. Wildbook: Crowdsourcing, computer vision, and data science for conservation. doi:10.48550/ARXIV.1710.08880

[5] Jeremy R. Brammer, Nicolas D. Brunet, A. Cole Burton, Alain Cuerrier, Finn Danielsen, Kanwaljeet Dewan, Thora Martina Herrmann, Micha V. Jackson, Rod Kennett, Guillaume Larocque, Monica Mulrennan, Arun Kumar Pratihast, Marie Saint-Arnaud, Colin Scott, and Murray M. Humphries. 2016. The role of digital data entry in participatory environmental monitoring. *Conservation Biology* 30, 6 (2016), 1277–1287. doi:10.1111/cobi.12727 arXiv:https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/cobi.12727

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification. arXiv:2406.08660 [cs.CL] https://arxiv.org/abs/2406.08660

[8] Jenna Burrell and Kentaro Toyama. 2009. What constitutes good ICTD research? *Information Technologies & International Development* 5, 3 (2009), pp–82.

[9] David Y. Cao, Jamie R. Silkey, Michael C. Decker, and Karolyn A. Wanat. 2024. Artificial intelligence-driven digital scribes in clinical documentation: Pilot study assessing the impact on dermatologist workflow and patient encounters. *JAAD International* 15 (June 2024), 149–151. doi:10.1016/j.jdin.2024.02.009

[10] Mustafa Chasmai, Wuao Liu, Subhransu Maji, and Grant Van Horn. 2025. Audio Geolocation: A Natural Sounds Benchmark. arXiv:2505.18726 [cs.SD] https://arxiv.org/abs/2505.18726

[11] Andrea Cuadra, Justine Breuch, Samantha Estrada, David Ihim, Isabelle Hung, Derek Askaryar, Marwan Hassanien, Kristen L. Fessele, and James A. Landay. 2024. Digital Forms for All: A Holistic Multimodal Large Language Model Agent for Health Data Entry. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2, Article 72 (May 2024), 39 pages. doi:10.1145/3659624

[12] Fred D Davis. 1989. Perceived Usefulness, Perceived Ease of Use and User Acceptance of Information Technology. *MIS quarterly* (1989).

[13] Dan Fitton, Keith Cheverst, Mark Rouncefield, Alan Dix, and Andy Crabtree. 2004. Probing technology with technology probes. In *Equator Workshop on Record and Replay Technologies*.

[14] Microsoft Cloud for Healthcare. 2025. Microsoft Dragon Copilot. https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot. [Accessed 05-08-2025].

[15] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open LLM Leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

[16] William Gaver, Anthony Dunne, and Elena Pacenti. 1999. Cultural Probes. *Interactions* VI (01 1999), 21–29.

[17] Georgi Gerganov. 2025. whisper.cpp: Port of OpenAI's Whisper model in C/C++. https://github.com/ggerganov/whisper.cpp.

[18] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim,

Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[19] Google Meet Help. 2025. Take notes for me in Google Meet. https://support.google.com/meet/answer/14754931?hl=en&ref_topic=14073938&sjid=8116723683457863554-NC. [Accessed 05-08-2025].

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

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

[22] iNaturalist contributors. 2025. iNaturalist Research-grade Observations. doi:10.15468/AB3S5X

[23] Chandra Irugalbandara, Ashish Mahendra, Roland Daynauth, Tharuka Kasthuri Arachchige, Jayanaka Dantanarayana, Krisztian Flautner, Lingjia Tang, Yiping Kang, and Jason Mars. 2024. Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's LLM with Open Source SLMs in Production. arXiv:2312.14972 [cs.SE] https://arxiv.org/abs/2312.14972

[24] Forest Isbell, Patricia Balvanera, Akira S Mori, Jin-Sheng He, James M Bullock, Ganga Ram Regmi, Eric W Seabloom, Simon Ferrier, Osvaldo E Sala, Nathaly R Guerrero-Ramírez, Julia Tavella, Daniel J Larkin, Bernhard Schmid, Charlotte L Outhwaite, Pairot Pramual, Elizabeth T Borer, Michel Loreau, Taiwo Crossby Omotoriogun, David O Obura, Maggie Anderson, Cristina Portales-Reyes, Kevin Kirkman, Pablo M Vergara, Adam Thomas Clark, Kimberly J Komatsu, Owen L Petchey, Sarah R Weiskopf, Laura J Williams, Scott L Collins, Nico Eisenhauer, Christopher H Trisos, Delphine Renard, Alexandra J Wright, Poonam Tripathi, Jane Cowles, Jarrett EK Byrnes, Peter B Reich, Andy Purvis, Zati Sharip, Mary I O'Connor, Clare E Kazanski, Nick M Haddad, Eulogio H Soto, Laura E Dee, Sandra Díaz, Chad R Zirbel, Meghan L Avolio, Shaopeng Wang, Zhiyuan Ma, Jingjing Liang, Hanan C Farah, Justin Andrew Johnson, Brian W Miller, Yann Hautier, Melinda D Smith, Johannes MH Knops, Bonnie JE Myers, Zuzana V Harmáčková, Jorge Cortés, Michael BJ Harfoot, Andrew Gonzalez, Tim Newbold, Jacqueline Oehri, Marina Mazón, Cynnamon Dobbs, and Meredith S Palmer. 2023. Expert perspectives on global biodiversity loss and its drivers and impacts on people. *Frontiers in Ecology and the Environment* 21, 2 (2023), 94–103. doi:10.1002/fee.2536 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.2536

[25] Roland Kays, Sarah C. Davidson, Matthias Berger, Gil Bohrer, Wolfgang Fiedler, Andrea Flack, Julian Hirt, Clemens Hahn, Dominik Gauggel, Benedict Russell, Andrea Kölzsch, Ashley Lohr, Jesko Partecke, Michael Quetting, Kamran Safi, Anne Scharf, Gabriel Schneider, Ilona Lang, Friedrich Schaeuffelhut, Matthias Landwehr, Martin Storhas, Louis van Schalkwyk, Candace Vinciguerra, Rolf Weinzierl, and Martin Wikelski. 2021. The Movebank system for studying global animal movement and demography. *Methods in Ecology and Evolution* 13, 2 (Dec. 2021), 419–431. doi:10.1111/2041-210x.13767

[26] Philip J Kroth, Nancy Morioka-Douglas, Sharry Veres, Katherine Pollock, Stewart Babbott, Sara Poplau, Katherine Corrigan, and Mark Linzer. 2018. The electronic elephant in the room: physicians and the electronic health record. *JAMIA open* 1, 1 (2018), 49–56.

[27] José J Lahoz-Monfort and Michael JL Magrath. 2021. A comprehensive overview of technologies for species and habitat monitoring and conservation. *BioScience* 71, 10 (2021), 1038–1062.

[28] Louis Liebenberg, Justin Steventon, !Nate Brahman, Karel Benadie, James Minye, Horekhwe (Karoha) Langwane, and Quashe (/Uase) Xhukwe. 2017. Smartphone Icon User Interface design for non-literate trackers and its implications for an inclusive citizen science. *Biological Conservation* 208 (April 2017), 155–162. doi:10.1016/j.biocon.2016.04.033

[29] Tsai-Ling Liu, Timothy C. Hetherington, Ajay Dharod, Tracey Carroll, Richa Bundy, Hieu Nguyen, Henry E. Bundy, McKenzie Isreal, Andrew McWilliams, and Jeffrey A. Cleveland. 2024. Does AI-Powered Clinical Documentation Enhance Clinician Efficiency? A Longitudinal Study. *NEJM AI* 1, 12 (Nov. 2024). doi:10.1056/aioa2400659

[30] Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv:2503.01743 [cs.CL] https://arxiv.org/abs/2503.01743

[31] William Moreto Mike Belecky, Rohit Singh. 2019. Life on the Frontline 2019: A Global Survey of the Working Conditions of Rangers.

[32] Birger Moëll and Fredrik Sand Aronsson. 2025. Journaling with large language models: a novel UX paradigm for AI-driven personal health management. *Frontiers in Artificial Intelligence* Volume 8 - 2025 (2025). doi:10.3389/frai.2025.1567580

[33] Chris Muashekele, Heike Winschiers-Theophilus, and Gereon Koch Kapuire. 2021. Integrating a community-based co-designed wildlife activity recording tool into a multi-stakeholder conservation management system. In *Proceedings of the 3rd African Human-Computer Interaction Conference: Inclusiveness and Empowerment* (Maputo, Mozambique) *(AfriCHI '21)*. Association for Computing Machinery, New York, NY, USA, 136–140. doi:10.1145/3448696.3448708

[34] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. *Survey Research in HCI*. Springer New York, 229–266. doi:10.1007/978-1-4939-0378-8_10

[35] Kathleen Musante and Billie R DeWalt. 2010. *Participant observation: A guide for fieldworkers*. Bloomsbury Publishing PLC.

[36] Khanyisile Ngcobo. 2025. Elon Musk's Starlink and the racially charged row over operating in South Africa — bbc.com. https://www.bbc.com/news/articles/cly3d8gd8mno.

[37] Rod Nickel Nqobile Dludla. 2024. Musk's Starlink ordered to cease operations in Namibia. https://www.reuters.com/technology/musks-starlink-ordered-cease-operations-namibia-2024-11-28/.

[38] Cornell Lab of Ornithology. 2021. eBird: An online database of bird distribution and abundance [web application]. http://www.ebird.org

[39] World Health Organization. 2025. Health workforce. https://www.who.int/health-topics/health-workforce. [Accessed 05-08-2025].

[40] Etosha National Park. [n. d.]. Etosha Climate — etoshanationalpark.co.za. https://etoshanationalpark.co.za/etosha-travel-tips/etosha-climate/.

[41] Stuart L Pimm, Sky Alibhai, Richard Bergl, Alex Dehgan, Chandra Giri, Zoë Jewell, Lucas Joppa, Roland Kays, and Scott Loarie. 2015. Emerging technologies to conserve biodiversity. *Trends in ecology & evolution* 30, 11 (2015), 685–696.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. doi:10.48550/ARXIV.2212.04356

[43] Domoina J. Rakotobe and Nancy J. Stevens. 2024. Closing staffing gaps in Madagascar's protected areas to achieve the 30 by 30 conservation target. *Conservation Science and Practice* 6, 5 (April 2024). doi:10.1111/csp2.13118

[44] Kirollos Risk. 2025. Fuse.js: Lightweight fuzzy-search, in JavaScript. https://www.fusejs.io/.

[45] James E. Siegler, Neha N. Patel, and C. Jessica Dine. 2015. Prioritizing Paperwork Over Patient Care: Why Can't We Do Both? *Journal of Graduate Medical Education* 7, 1 (March 2015), 16–18. doi:10.4300/jgme-d-14-00494.1

[46] Aaron A Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery* 5, 3 (2024), CAT–23.

[47] Aaron A. Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Sarah B. Wilson Hannay, Kate Castilla, Cindy S. Lau, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2025. Ambient Artificial Intelligence Scribes: Learnings after 1 Year and over 2.5 Million Uses. *NEJM Catalyst* 6, 5 (April 2025). doi:10.1056/cat.25.0040

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[49] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.

[50] Anthony Waldron, Daniel C Miller, Dave Redding, Arne Mooers, Tyler S Kuhn, Nate Nibbelink, J Timmons Roberts, Joseph A Tobias, and John L Gittleman. 2017. Reductions in global biodiversity loss predicted from conservation spending. *Nature* 551, 7680 (2017), 364–367.

[51] Jake Wall, Jes Lefcourt, Chris Jones, Chris Doehring, Dan O'Neill, Dennis Schneider, Jordan Steward, Joshua Krautwurst, Tiffany Wong, Bruce Jones, Karen Goodfellow, Ted Schmitt, Kathleen Gobush, Iain Douglas-Hamilton, Frank Pope, Eric Schmidt, Jonathan Palmer, Emma Stokes, Andrea Reid, L. Mark Elbroch, Peter Kulits, Catherine Villeneuve, Victor Matsanza, Geoff Clinning, Jordi van Oort, Kristen Denninger Snyder, Alina Peter Daati, Wesley Gold, Stephen Cunliffe, Batian Craig, Barry Cork, Grant Burden, Marc Goss, Nathan Hahn, Sarah Carroll, Eric Gitonga, Ray Rao, Jared A. Stabach, Frédéric Dulude-de Broin, Patrick Omondi, and George Wittemyer. 2024. EarthRanger: An open-source platform for ecosystem monitoring, research and management. *Methods in Ecology and Evolution* 15, 11 (2024), 1968–1979. doi:10.1111/2041-210X.14399 arXiv:https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.14399

[52] Jake Wall, George Wittemyer, Brian Klinkenberg, and Iain Douglas-Hamilton. 2014. Novel opportunities for wildlife conservation and research with real-time monitoring. *Ecological Applications* 24, 4 (2014), 593–601. doi:10.1890/13-1971.1 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/13-1971.1

[53] Richmond Y. Wong, Vera Khovanskaya, Sarah E. Fox, Nick Merrill, and Phoebe Sengers. 2020. Infrastructural Speculations: Tactics for Designing and Interrogating Lifeworlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, 1–15. doi:10.1145/3313831.3376515

[54] Susan Wyche. 2022. Reimagining the Mobile Phone: Investigating Speculative Approaches to Design in Human-Computer Interaction for Development (HCI4D). *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–27. doi:10.1145/3555648

[55] Xiao Yan and Yi Ding. 2025. Are We There Yet? A Measurement Study of Efficiency for LLM Applications on Mobile Devices. arXiv:2504.00002 [cs.PF] https://arxiv.org/abs/2504.00002

[56] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[57] Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. LoRA Land: 310 Fine-tuned LLMs that Rival GPT-4, A Technical Report. arXiv:2405.00732 [cs.CL] https://arxiv.org/abs/2405.00732

A Appendix

A.1 Survey Questions

1. What organization do you work with? [Free text]
2. What language do you use EarthRanger Event forms in, and is it your native language? If not, what is your native language? [Language Dropdown]
3. How often do you enter a datapoint or observations through the EarthRanger Event Form?
    - Not at all
    - A few times a year
    - A few times a month
    - Weekly
    - Daily
4. About how long have you been using EarthRanger Event forms?
    - 1-2 years
    - 2-5 years
    - 5 years or more
5. How often do you work with the EarthRanger software team to update or add forms?
    - Monthly
    - Yearly
    - Some other cadence [Free text]
6. Please select all of the following conditions in which you use the EarthRanger Event Form.
    - Outdoors
    - In an office setting
    - On a mobile phone
    - Without internet connection
    - In the dark / at night
    - Under bright sunlight
    - With loud background noise
    - In dangerous conditions
    - While operating a vehicle
    - While hiking or running
    - When you have to be quiet
    - In bad weather
7. How satisfied are you with the functionality of EarthRanger Event forms?
    - Extremely dissatisfied
    - Somewhat dissatisfied
    - Neither satisfied nor dissatisfied
    - Somewhat satisfied
    - Extremely Satisfied
8. Please elaborate briefly on your answer to the previous question — what do you like, or what would you like to see change? [Free text]
9. What did you use before the EarthRanger event form for in-the-field data collection, if anything? [Free text]
10. Do you ever find data entry with EarthRanger Event Forms slow or difficult? Please elaborate briefly. [Free text]
11. How would you feel about using your voice to fill out Event Forms, instead of typing? [Free text]

A.2 Maintenance Playbook

Here, we enumerate the hardware, personnel, and processes required to maintain our system over time.

Hardware. The main hardware requirement for fine-tuning of the specific model utilized (Llama 3.2-1B) is, realistically, a GPU with at least 4GB of VRAM. While training can be conducted on a CPU, the process will take much longer than on GPU. Larger models may require more VRAM — enough space for the model weights themselves, then additional space for LORA module activations.

Data. If the organization lacks the staff or time to collect audio-form pairs (the gold standard for fine-tuning data), then the next-best alternative is to synthesize data. We did this by randomly sampling from all form fields and options, then recombining these fields in natural language formulaically. Using even a large LLM to generate natural-sounding language led to hallucinations, with a negative effect on downstream fine-tuning, so we restricted the use of LLMs to generating short sentences for any open-response form fields. This dataset of sampled form field options and natural language observations can then be used directly for model tuning.

Machine Learning (ML) Maintainer. This individual is responsible for the three major steps of machine learning deployment: (1) training, (2) integration, and (3) monitoring.

1. **Training**: From our current workflow, this step consists of generating a realistic synthetic dataset covering the span of all field options across form types, creating prompts for each form type, and training the language model to an acceptable accuracy. This task also includes the prompting of the audio model.
2. **Integration**: After training is complete, the model must be converted to a CPU-efficient format, wrapped in an interface which uses regex, fuzzy search, and type-checking to restrict hallucinations and other output errors, and integrated with the EarthRanger application.
3. **Monitoring**: After deployment, the ML maintainer must ensure that the model does not lose accuracy over time by soliciting feedback from users and monitoring the model outputs themselves. Here, the ML maintainer must also ensure that any new users' accents and language dialects are accurately transcribed by the audio model.

Organization-Level EarthRanger Super User. The super user is often a technical staff member, who may also be responsible for data analysis and coordinating hardware, for example. This person is responsible for customizing EarthRanger forms, along with their organizational role more broadly. Upon receiving or establishing a new research target, this person would, together with field staff, evaluate what new forms or fields should be added to the existing set of forms.

Field Staff. The field staff play a critical feedback role in the maintenance of the system: they are the ones who test systems in the field, and they also provide feedback on how completely and effectively the forms capture data in the field. After the addition of the audio and language models, the field staff are also the first to notice if either model is failing.